The New Encyclopædia Britannica

Volume 20

MACROPÆDIA

Knowledge in Depth

FOUNDED 1768
15 TH EDITION

Encyclopædia Britannica, Inc.
Robert P. Gwinn, Chairman, Board of Directors
Peter B. Norton, President
Philip W. Goetz, Editor in Chief

Chicago
Auckland/Geneva/London/Madrid/Manila/Paris
Rome/Seoul/Sydney/Tokyo/Toronto



THE UNIVERSITY OF CHICAGO

"Let knowledge grow from more to more
and thus be human life enriched."

The *Encyclopædia Britannica* is published with the editorial advice of the faculties of the University of Chicago.

Additional advice is given by committees of members drawn from the faculties of the Australian National University, the universities of British Columbia (Can.), Cambridge (Eng.), Copenhagen (Den.), Edinburgh (Scot.), Florence (Italy), Leiden (Neth.), London (Eng.), Marburg (Ger.), Montreal (Can.), Oxford (Eng.), the Ruhr (Ger.), Sussex (Eng.), Toronto (Can.), Victoria (Can.), and Waterloo (Can.); the Complutensian University of Madrid (Spain); the Max Planck Institute for Biophysical Chemistry (Ger.); the New University of Lisbon (Port.); the School of Higher Studies in Social Sciences (Fr.); Simon Fraser University (Can.); and York University (Can.).

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

Eleventh Edition
© 1911
By Encyclopædia Britannica, Inc.

Twelfth Edition
© 1922
By Encyclopædia Britannica, Inc.

Thirteenth Edition
© 1926
By Encyclopædia Britannica, Inc.

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

Fifteenth Edition
© 1974, 1975, 1976, 1977, 1978, 1979, 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989, 1990, 1991
By Encyclopædia Britannica, Inc.

© 1991
By Encyclopædia Britannica, Inc.

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

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

Printed in U.S.A.

Library of Congress Catalog Card Number: 89-81675
International Standard Book Number: 0-85229-529-4

CONTENTS

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

Geomorphic Processes

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

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

This article is divided into the following sections:

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

Physiographic effects of tectonism

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

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

RECENT TECTONIC MOVEMENTS

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

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

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

Mountain regions of intensive seismic activity

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

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

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

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

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

EFFECTS ON THE EARTH'S SURFACE

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

Rates of continental denudation

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

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

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

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

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

Mudflows as an example of natural denudation

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

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

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

It has been suggested that all tectonic movements that are expressed in present-day relief should be called neotectonics. In papers by Soviet scientists this connotation

Neotectonics as the basis of present-day relief

emerges from considerations of the general amplitude and character of the most recent tectonic movements. A review map of these movements in the Soviet Union served as a model for international compilations of this type. The main basis for the work was study of the thickness of Pliocene–Quaternary deposits (younger than 7,000,000 years) that accumulated in tectonic subsidence areas (depressions and basins) and of the amplitude of uplift of ancient surfaces in mountain ridges. In these instances, the obtained values of uplift and deposition amounted to hundreds of metres of relief over distances of several kilometres. The role of neotectonics in forming the present relief features is acknowledged by all authorities. (I.P.G.)

Weathering

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

Weathering and landforms

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

PROCESSES INVOLVED IN WEATHERING

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

Thermal expansion and contraction

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

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

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

Pressure-release hypothesis

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

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

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

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

Freeze–thaw cycles and effects

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

Crystallization of such salts as sodium chloride and gypsum is also cited as a cause of rock disintegration, particularly in arid regions. There seems little doubt that

Crystal-growth pressures

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

The role of plant growth and organisms

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

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

Solution effects

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

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

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

Hydration and hydrolysis

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

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

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

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

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

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

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

CONTROLS AND RATES OF WEATHERING

Several factors control the type and rate of rock weathering.

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

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

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

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

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

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

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

Rates of weathering

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

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

Slope movements

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

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

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

Significance to man

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

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

FACTORS PRODUCING SLOPE MOVEMENTS

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

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

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

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

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

Effects of earthquakes, groundwater, and climate

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

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

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

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

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

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

TYPES OF SLOPE MOVEMENTS

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

Origin and occurrence of rockfalls

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

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

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

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

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

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

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

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

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

Deformation by bulging

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

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

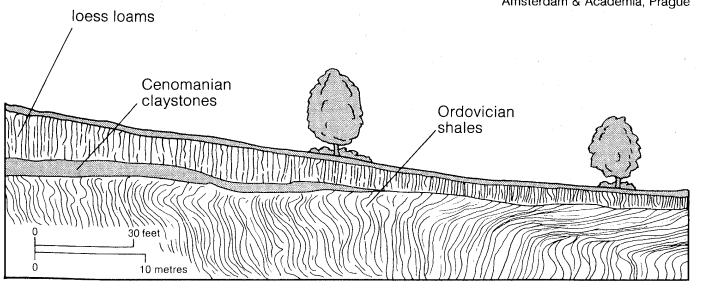


Figure 1: Creep and bending of beds in a loam pit at Prague, Czechoslovakia (see text).

In some regions the squeezing out of soft rocks in the lower parts of the valley slopes is so widespread that it causes serious economic problems. Bulging was first described from an area of iron-ore mining near Northampton (central England), where valley sides are composed of solid Jurassic limestones and shales (136,000,000 to 190,000,000 years old) that overlie soft Lias clays. Although the beds are nearly horizontal, the near-surface beds of solid rocks are inclined into the slopes, whereas the clays at the foot of the slopes are squeezed upward and contorted. In the initial stage, the bulging appears as a slight anticlinal bend of beds; with advancing deformation the clays are bent into folds, and even small faults may form at the foot of the slope. Although the disturbances extend to several tens of metres in depth, from the geological point of view they are surficial phenomena because they disappear both downward and into the slope where the beds preserve a normal subhorizontal course. This deformation can be interpreted in terms of the upward squeezing of plastic substance from the loaded medium into the unloaded one. The stress that caused the heave of clays results from the difference in the loading of clays in the bottom of the valley and under its slopes.

Figure 2 shows an example of bulging in a river valley near Ostrava (Czechoslovakia). The valley occurs in marly shales of Cretaceous age (65,000,000 to 136,000,000 years old) that are pierced by sills of very firm volcanic rock called teschenite. The section shows that the teschenite body is broken into several blocks by a system of faults running roughly parallel to the valley. The soft shales that are squeezed out by the heavy rock blocks move toward the stream, which gradually carries them away. The main deformations probably occurred about 230,000 years ago, during the Pleistocene Epoch (10,000 to 2,500,000 years ago), because the steps between the blocks are filled with slope debris and loess loam.

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

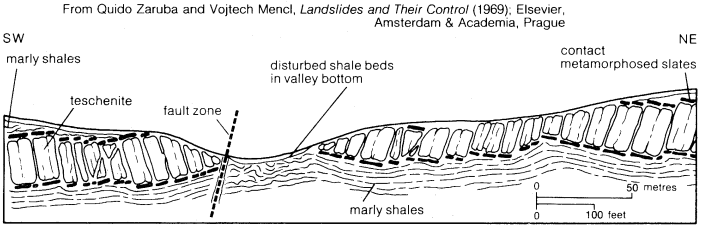


Figure 2: Squeezing-out of marly shales on the valley bottom of the Lucina River, Czechoslovakia.

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

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

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

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

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

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

Slide characteristics

The sliding movements involve bedrock or surficial deposits, but very often both bedrock and its cover are involved. The part of the slope that moves separates from the remaining mass along the plane of least resistance (Figure 3). This slip surface is commonly formed by a

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

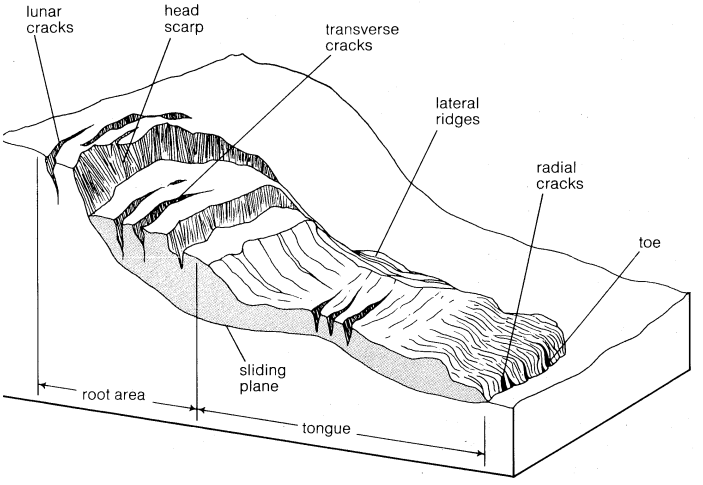


Figure 3: Principal parts of a landslide and characteristic cracks.

bedding, joint, or fault plane. The movements are generally rapid and originate when the planes of separation dip downslope and their continuity is disturbed at the foot of the slope. In stratified rocks with smooth, even bedding planes, the dip of beds is usually the maximum inclination at which the slope is permanently stable. If the beds are undercut by stream erosion, they maintain their position only by friction, which increases with the roughness and unevenness of the bedding planes. Friction can be reduced by climatic factors: freezing and thawing of interstitial water, or by hydrostatic pressure of water in the joints if the free outflow of water is obstructed. Failure can also be provoked by the increase of slope angle, as a result of uplift of the Earth's crust.

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

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

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

Form and occurrence of slumps

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

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

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

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

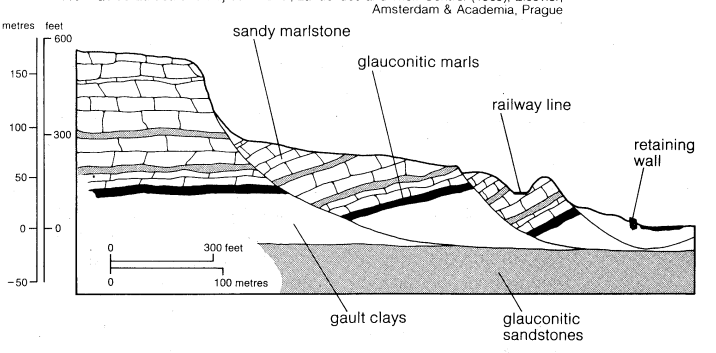


Figure 4: Landslide on the seashore near Folkestone, Kent.

Role of water saturation

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

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

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

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

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

Failure of quick clays

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

CHARACTERISTICS OF UNSTABLE SLOPES

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

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

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

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

Fluvial processes

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

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

ENTRAINMENT AND TRANSPORT OF SEDIMENTARY PARTICLES

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

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

Modes of sediment transport

Particle size and settling velocity

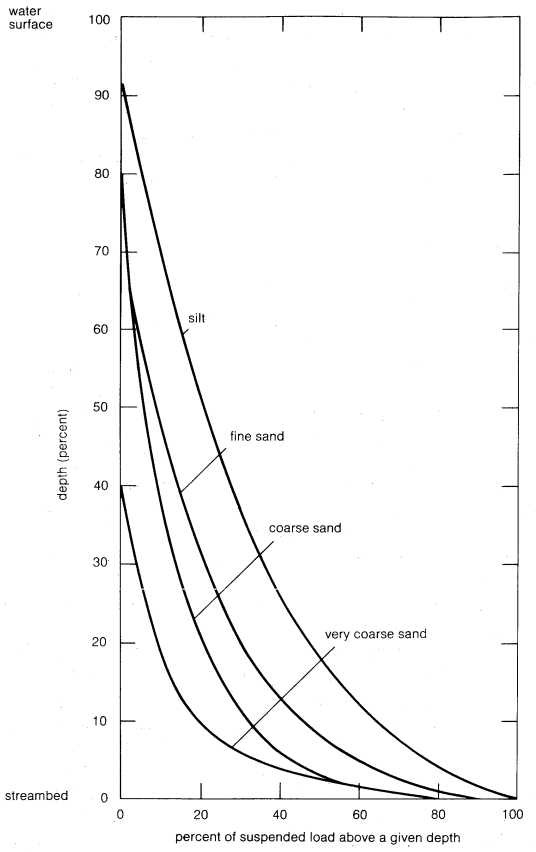


Figure 5: *Distribution of suspended load with depth for particles of different sizes.*
Coarse particles are concentrated closer to the bed, while finer particles, such as silt, are more uniformly distributed with depth.

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

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

MATERIALS TRANSPORTED BY NATURAL RIVERS

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

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

EROSION AND DEPOSITION IN NATURAL CHANNELS

Equilibrium conditions and the migration of meanders

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

Dissolved and Suspended Load in Selected Rivers in Different Climatic Regions

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

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

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

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

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

Flow and transport variation with time

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

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

From *Bulletin of the U.S. Geological Survey* (1960)

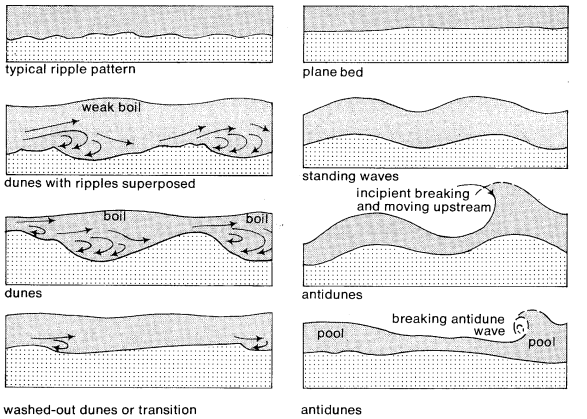


Figure 6: *Forms of bed roughness in alluvial channels.* This sequence of forms is related to the depth and velocity of flow and to the rate of sediment transport.

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

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

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

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

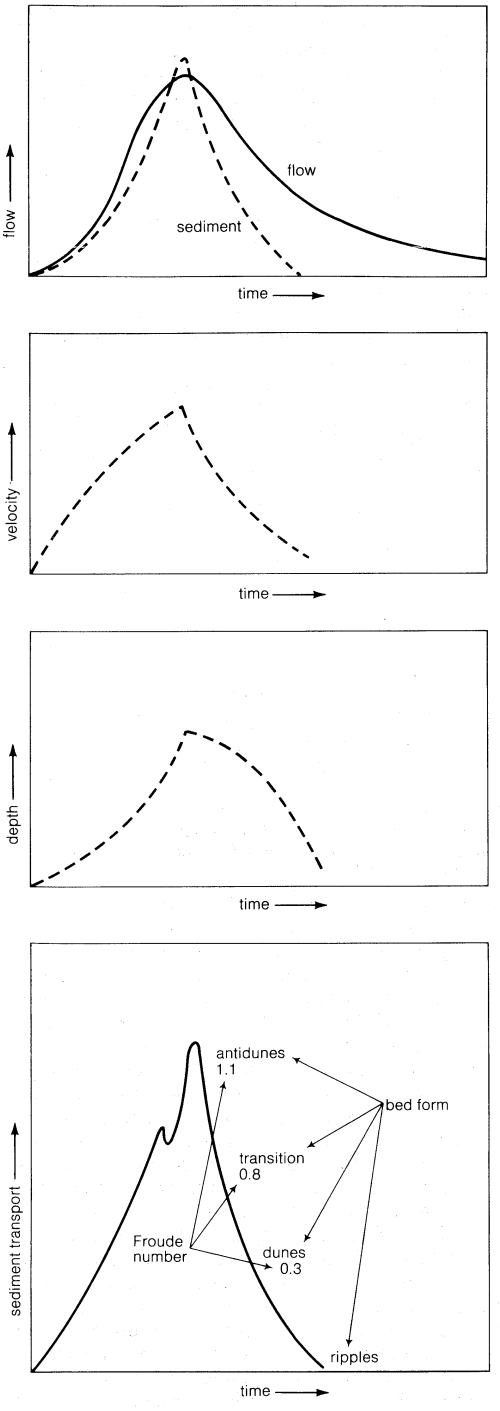


Figure 7: Sequential changes in flow (discharge), velocity, depth, and sediment transport associated with the passage of a flood in a natural channel with a sand bed. Changes in bed form and the approximate Froude number at which such changes might be observed are shown in association with changes in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

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

Magnitude and frequency of flow events

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

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

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

Changes in hydrologic conditions

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

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

Rates of erosion and deposition

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

FLUVIAL PROCESSES IN DIFFERENT ENVIRONMENTS

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

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

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

Grant Heilman

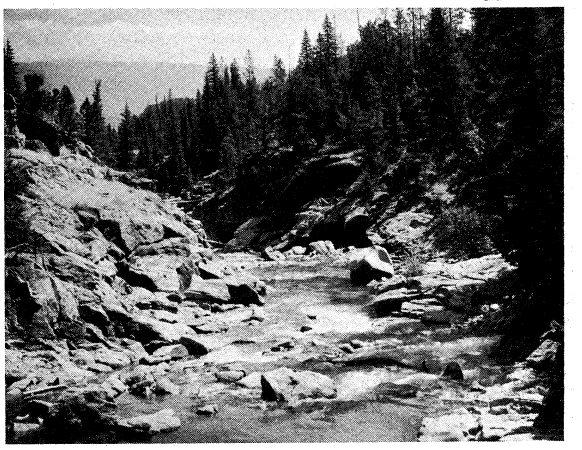


Figure 8: A boulder-bed high mountain stream, Crandall Creek, Wyoming.

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

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

Dry washes

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

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

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

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

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

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

Waterfalls

Josef Muench

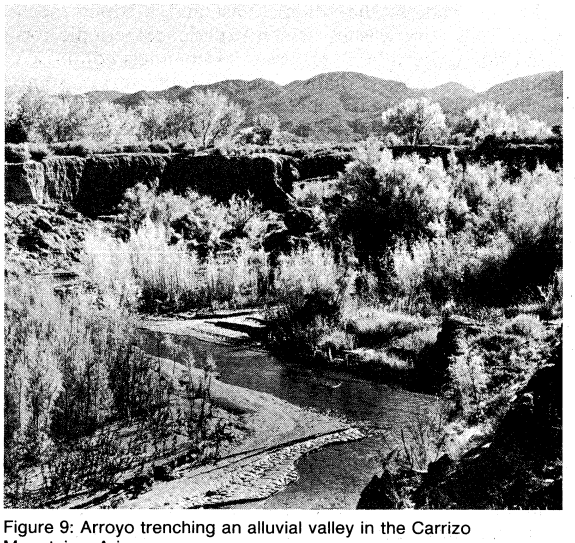


Figure 9: Arroyo trenching an alluvial valley in the Carrizo Mountains, Arizona.

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

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

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

Processes of glaciation

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

GLACIAL EROSION

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

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

Glacial quarrying

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

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

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

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

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

GLACIAL TRANSPORT

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

Ablation and accumulation of glaciers

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

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

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

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

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

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

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

GLACIAL DEPOSITION

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

Fluvioglacial processes at the ice margin

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

These conditions give rise to a great variety of glacially

By courtesy of James Zumberge, University of Arizona, Tucson



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

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

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

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

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

GLACIAL LOADING AND UNLOADING

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

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

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

Postglacial uplift in progress today

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

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

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

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

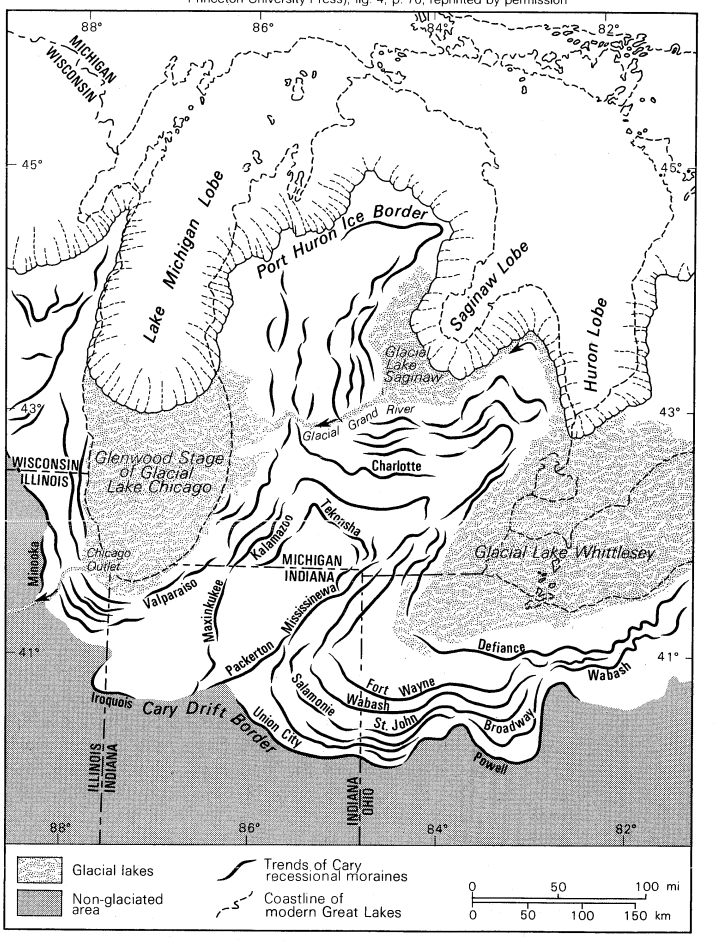


Figure 11: Upper Great Lakes region during Port Huron maximum, about 13,000 years ago.

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

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

PERIGLACIAL PROCESSES

Periglacial environment

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

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

Wind action

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

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

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

TRANSPORTATION OF ROCK DEBRIS BY WIND

Turbulent flow and the boundary layer

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

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

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

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

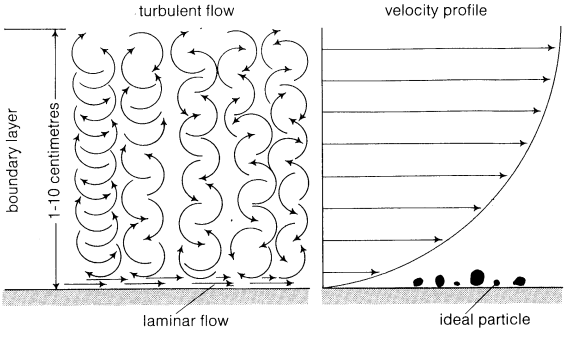


Figure 12: Diagrammatic view of the laminar and turbulent flow fields and the velocity distribution of the wind above the ground surface. The smallest particles are not lifted because they do not project above the laminar flow zone. The largest particles are subject to turbulent lifting forces but are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

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

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

Saltation of grains

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

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

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

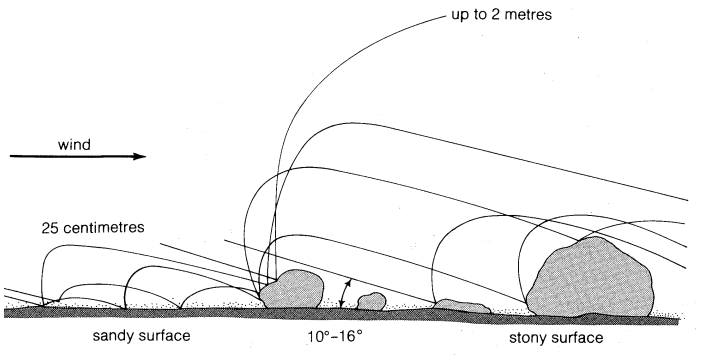


Figure 13: Trajectories and general distribution of particles transported by wind over (left) sandy surface and (right) stony surface. Note the much higher paths of particles on the stony surface, which result from impact with larger particles.

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

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

Impact creep

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

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

Movement of large particles

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

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

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

Intercontinental dust distribution

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

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

EFFECTS OF WIND TRANSPORT

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

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

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

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

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

Production of desert pavement

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

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

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

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

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

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

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

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

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

The maximum rate of wear by sandblasting in natural situations occurs at a discrete height above ground level. This probably reflects the greater impact energy of saltating grains travelling at that height. The shape of most sandblasted faces and specific field experiments suggest that maximum wear occurs mostly between five and 25

By courtesy of Robert P. Sharp

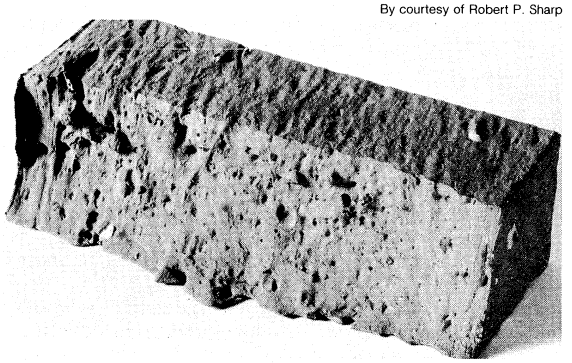


Figure 14: Common red brick after six years of natural sandblasting.

centimetres (two and 10 inches) above the ground. At one site, subject to frequent sand saltation over a bouldery surface, maximum wear over an 11-year interval occurred at a height of 22 centimetres (nine inches).

Ventifacts, flutes, and grooves

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

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

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

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

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

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

Conditions favourable to sandblasting

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

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

DEPOSITION BY WIND

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

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

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

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

Growth and structure of dunes

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

As the mound gets higher, it develops a slight asymmetry with a gentler windward slope and a somewhat steeper leeward slope. Under unusually strong winds, separation of the boundary layer occurs at the dune crest, resulting in exceptional deposition of traction and saltation sand just downwind. This creates a steep lee face inclined at the angle of repose for loose sand, about 30°; at greater angles material will slide. The mound now becomes an efficient sand trap; it is truly a dune. The more sand it traps, the higher the lee face grows and the more efficient is the trapping mechanism. Sand travelling by traction and creep is dumped over the brink of the lee face, and saltating grains rain down upon it from the separated boundary layer (Figure 15). Deposition is greater on the upper reach of the lee face, causing it to become increasingly steeper. Under such conditions the slope steepens to an inclination of

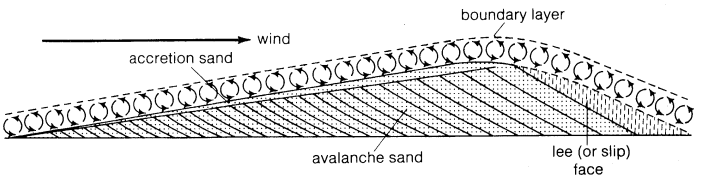


Figure 15: Distribution of sand on and within a dune due to wind action. Sand is shown raining down onto the lee or slip face; this general process leads to dune migration.

about 34° before it gives way by avalanching to establish a more normal angle of repose (30° to 33°). For this reason, the lee slope has been called the slip face. The lee slope of an active dune is constantly being scarred by little sand avalanches all along its face. Similar avalanches can be initiated artificially by a person walking along the brink of a freshly built lee slope. The flowing sand generates a very low-pitched noise, often heard in dune areas when strong winds are blowing and active deposition is underway.

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

Migrating dunes advance by deposition of material on the lee face. In time, therefore, under ideal circumstances they must consist primarily of layers of sand laid down near the angle of repose (30°–33°) inclined in the direction of dune advance. In many dunes the internal bedded structure departs from this simple model, but, nonetheless, it has proved a useful concept in the determination of paleo-wind directions indicated by ancient deposits of dune sand.

Surfaces of loose sand across which wind is transporting sand usually develop a rippled configuration. The ripples are small, essentially parallel, asymmetrical ridges composed largely of sand moving by traction and impact creep, although some exchange with the saltating curtain

Wind ripples

By courtesy of Robert P. Sharp

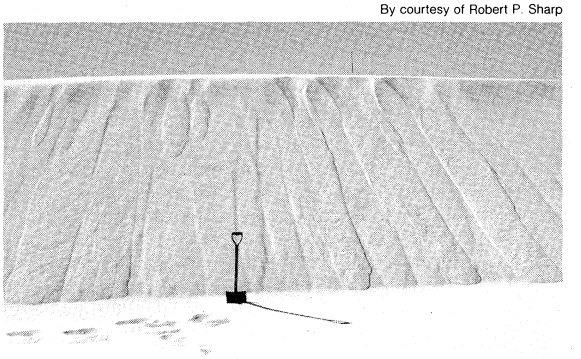


Figure 16: Sand avalanche tongues on lee face of a recently active dune.

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

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

Role of vegetation. Windblown sand and vegetation display some significant interrelationships. Vegetation of almost any type inhibits transport by traction, creep, and saltation. The resulting deposition can overwhelm and kill the vegetation, although some vegetation can survive partial or even complete burial by a migrating dune and

By courtesy of Robert P. Sharp

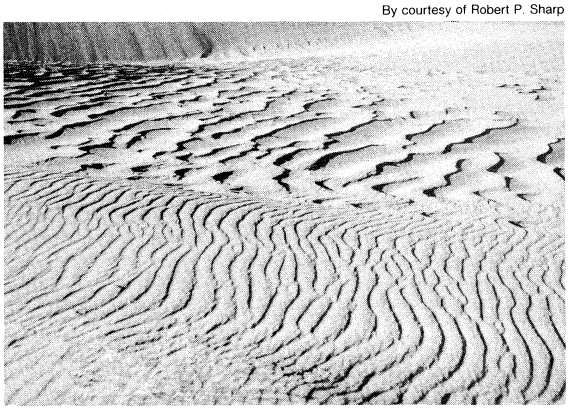


Figure 17: Wind ripples formed in sand (foreground) and in sand and fine gravel (midground).

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

Deposition of coastal dunes

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

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

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

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

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

Physiographic effects of man

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

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

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

RISE OF HUMAN AGENCY

Importance of fire

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

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

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

Ancient irrigation works

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

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

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

Materials for the pyramids and other works

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

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

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

EFFECTS OF MODERN MAN

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

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

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

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

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

Quarries and excavations

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

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

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

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

Harbours, breakwaters, and dikes

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

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

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

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

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

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

Underground fluids

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

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

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

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

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

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

Urbanization and hydrologic change

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

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

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

Soil erosion

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

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

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

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

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

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

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

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

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

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

BIBLIOGRAPHY

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

(Ed.)

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

(I.P.G./Ed.)

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

(C.R.T.)

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

(Q.Z./Ed.)

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

(M.G.W.)

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

(J.H.Z.)

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

(R.P.Sp.)

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

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

German Literature

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

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

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

The article is divided into the following sections:

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

Old High German period

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

Heldenlieder

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

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

Pagan works

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

Middle High German period

CHARACTERISTICS OF THE PERIOD

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

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

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

The *Nibelungenlied*

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

PRINCIPAL LITERARY FORMS

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

The foremost poets of the court epic

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

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

The love lyrics of Walther von der Vogelweide

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

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

Morality and miracle plays

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

From Middle High German to Baroque

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

THE COURTLY REVIVAL AND THE HUMANIST MOVEMENT

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

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

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

The humanist impact was more direct in the drama.

Revitalization of drama by humanism

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

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

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

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

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

THE REFORMATION

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

Literature of the Reformation

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

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

Drama as a vehicle for Lutheran teaching

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

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

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

The works of Hans Sachs

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

THE CLOSE OF THE 16TH CENTURY

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

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

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

Source of the Faust story

The 17th century

THE CHARACTER OF THE PERIOD

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

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

MAJOR WORKS OF THE PERIOD

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

Martin Opitz' influence

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

Characteristics of Baroque poetry

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

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

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

The achievement of *Simplicissimus*

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

The 18th century

THE AGE OF ENLIGHTENMENT

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

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

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

The influence of English translations

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

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

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

Lessing's dramatic works

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

Wieland's contributions

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

THE AGE OF GOETHE

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

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

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

The influence of Herder

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

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

The dramas of young Schiller

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

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

Moral idealism

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

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

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

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

Goethe's *Faust*

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

THE ROMANTIC MOVEMENT

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

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

Hölderlin and Jean Paul

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

Wackenroder, Tieck, and Novalis

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

Interest in folk heritage

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

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

The Heidelberg school broke up about 1809, and Berlin and Dresden became the main centres of Romanticism. Two major writers are associated with this phase, Heinrich von Kleist and E.T.A. Hoffmann. Both explored the darker aspects of life that had always interested the Romantics. Kleist was a master of the *Novelle* form (*Erzählungen* [1810–11; "Tales"; Eng. trans. *The Marquise of O. and Other Stories*]) and the greatest dramatist of his day. His plays include *Amphitryon* (1807), *Penthesilea* (1808), and *Prinz Friedrich von Homburg* (posthumously published, 1821). In such works, Kleist depicted the tragic impossibility of understanding a world riddled with ambiguities. Hoffmann, a storyteller of genius, gave European currency to German Romanticism in countless works of fantasy and the grotesque. Among his tales are *Der goldne Topf* (1814; *The Golden Pot*) and *Das Fräulein von Scudéri* (1819; *Madame de Scudéri*). Other figures of note in this phase were the dramatist Zacharias Werner and the poet Wilhelm Müller, whose song cycles such as *Die Winterreise* (1824; "The Winter Journey") were set to music by Franz Schubert.

Kleist and Hoffmann

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

The 19th century

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

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

GRILLPARZER, BÜCHNER, AND THE DRAMA

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

Grillparzer's historical plays

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

Significance of Büchner's *Woyzeck*

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

LYRIC POETRY

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

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

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

Heine's attack on Romanticism

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

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

iterature
f socio-
olitical
ommit-
ıent

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

REALISM AND REGIONALISM

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

:haracteristics
f Poetic
ealism

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

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

The gentle art of Stifter

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

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

The underlying theme of Keller's works

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

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

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

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

Meyer's fusion of poetry and reality

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

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

The social novels of Fontane

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

Nietzsche's theory of art

NATURALISM

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

Hauptmann's plays

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

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

The 20th century

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

MAJOR LITERARY TRENDS AND CONDITIONS

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

George's Symbolist poetry

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

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

The end-of-the-century analyses of decadence

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

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

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

Rilke's view of God

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

Mann's analyses of his age

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

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

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

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

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

The nightmare vision of Franz Kafka

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

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

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

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

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

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

LITERATURE AFTER WORLD WAR II

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

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

Böll's writings

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

Grass's portrayal of modern German history

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

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

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

The plays of Bertolt Brecht

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

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

Documentary school of drama

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

The New German Radio Play

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

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

Trends in literary criticism

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

BIBLIOGRAPHY. MALCOLM PASLEY (ed.), *Germany: A Companion to German Studies,* 2nd ed. (1982), good background reading; HERBERT A. FRENZEL and ELISABETH FRENZEL, *Daten deutscher Dichtung: Chronologischer Abriss der deutschen Literaturgeschichte,* 20th ed., 2 vol. (1982), an indispensable chronology of authors, works, and movements; JOHN G. ROBERTSON, *A History of German Literature,* 6th ed. (1970), an essential reference work providing easy access to dates and facts; C.P. MAGILL, *German Literature* (1974); GILBERT WATERHOUSE, continued by H.M. WAIDSON, *A Short History of German Literature,* 3rd ed. rev. (1959), an excellent brief outline; and FRITZ MARTINI, *Deutsche Literaturgeschichte,* 18th ed. (1984).

Studies that focus on specific periods or trends of German literary history include J. KNIGHT BOSTOCK, *A Handbook on Old High German Literature,* 2nd ed. rev. by K.C. KING and D.R. MCLINTOCK (1976), an essential reference work; M. O'C. WALSHE, *Medieval German Literature* (1962), a very useful reference; PAUL SALMON, *Literature in Medieval Germany* (1967), an excellent introduction; ARCHER TAYLOR, *Problems in German Literary History of the Fifteenth and Sixteenth Centuries* (1939, reprinted 1966); RICHARD NEWALD, *Die deutsche Literatur vom Späthumanismus zur Empfindsamkeit, 1570–1750,* 6th ed. (1967); ROY PASCAL, *German Literature in the Sixteenth and Seventeenth Centuries* (1968, reprinted 1979), one of the best English introductions to this period; MARIAN SZYROCKI, *Die Deutsche Literatur des Barock* (1968, reissued 1979), a valuable source; ROBERT M. BROWNING, *German Baroque Poetry, 1618–1723* (1971); FRIEDHELM RADANT, *From Baroque to Storm and Stress, 1720–1775* (1977), a most useful survey; WALTER H. BRUFORD, *Germany in the Eighteenth Century: The Social Background of the Literary Revival* (1935, reprinted 1971); and *Culture and Society in Classical Weimar, 1775–1806* (1962), informative studies of the cultural milieu of the period; HENRY B. GARLAND, *Storm and Stress* (1952), an adequate introduction; WILLIAM D. ROBSON-SCOTT, *The Literary Background of the Gothic Revival in Germany: A Chapter in the History of Taste* (1965), on art as well as literature; T.R. REED, *The Classical Centre. Goethe and Weimar, 1772–1832* (1980), a masterly study; ERNST L. STAHL and W.E. YUILL, *German Literature of the Eighteenth and Nineteenth Centuries* (1970), a solid introductory study; GLYN TEGAI HUGHES, *Romantic German Literature* (1979), an excellent survey; GEORG BRANDES, *Main Currents in 19th-Century Literature,* vol. 6, *The Young Germany* (1901–05, reissued 1975; originally published in Danish, 1872–90), a thorough work on European comparative literature; RONALD GRAY, *The German Tradition in Literature, 1871–1945* (1965), an interesting study; JETHRO BITHELL, *Modern German Literature, 1880–1950,* 3rd ed. (1959, reprinted 1963), a useful reference work; AUGUST CLOSS (ed.), *Twentieth Century German Literature* (1969), an excellent introduction, one in the series of *Introductions to German Literature;* RAYMOND FURNESS, *The Twentieth Century, 1890–1945* (1978); WALTER H. SOKEL, *The Writer in Extremis: Expressionism in Twentieth-Century German Literature* (1959, reissued 1964).

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

Germany

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

Of historical, if no longer political, importance is the long internal boundary that for 40 years partitioned Germany into two nations. It was based on the lines of demarcation agreed upon at the Yalta Conference of 1945, separating the then-Soviet occupation zone of Germany from the zones occupied by the Western allies, on whose territory West Germany subsequently emerged. On the East German side, this boundary was, until the fall of the communist government in 1989, marked by defenses in depth designed to prevent the escape of the population. The 185 square miles (480 square kilometres) of the "island" of West Berlin were similarly ringed from 1961 to 1989 by the Berlin Wall. Although West Berlin's existence was secured by a 1971 agreement among the four wartime Allies, the city declined in national and international significance until the events of 1989–90 restored a united Berlin as the capital of a united Germany.

The Federal Republic has a maximum north-south extent of about 530 miles (850 kilometres), between latitudes 47° and 55° N, and an east-west maximum (across the middle of the country) of about 400 miles (640 kilometres), between longitudes 6° and 15° E. It has an area of 137,855 square miles (357,042 square kilometres).

The constitution of the Federal Republic, adopted in 1949, gives most government powers to its constituent *Länder* (states), which formerly numbered 11 (after three original states merged) but which grew to 16 (including Berlin) upon the accession of East Germany; only matters of undoubted importance for the nation as a whole, such as defense and foreign affairs, are reserved to the federal government. At both the *Land* and federal levels, parliamentary democracy prevails. The Federal Republic has been a member of the North Atlantic Treaty Organization (NATO) since 1955 and was a founding member of the European Coal and Steel Community in 1952 and of the European Economic Community in 1957. During the four decades of partition, the Federal Republic concluded a number of agreements with the Soviet Union and East Germany, which it supported to some extent economically in return for various concessions with regard to access to Berlin and humanitarian matters. West Germany's rapid economic recovery in the 1950s (*Wirtschaftswunder,* or "economic miracle") brought it into a leading position among the world's economic powers, a position that it has subsequently maintained. Reunion with the eastern sector, while constituting a huge financial burden at the outset, was expected only to strengthen this standing.

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

The largest of the *Länder* is Bavaria (German: Bayern), the most populous is North Rhine-Westphalia (Nordrhein-Westfalen), and the wealthiest is Baden-Württemberg. In the extreme north lies Schleswig-Holstein, south of which is Lower Saxony (Niedersachsen). The ancient free Hanseatic cities of Hamburg and Bremen rank as *Länder* in their own right. The states of Rhineland-Palatinate (Rheinland-Pfalz) and Hesse (Hessen) cover the central area of the Federal Republic; the mineral-rich Saarland (often called the Saar) is a pocket in the southwestern corner of the Pfälzer Mountains (Pfälzerwald). Five *Länder,* in addition to that of Berlin, were formed in 1990 from the territory of the former Democratic Republic: Mecklenburg-West Pomerania (Mecklenburg-Vorpommern) in the north, Brandenburg (surrounding Berlin) in the east, Saxony (Sachsen) in the southeast, Thuringia (Thüringen) in the southwest, and Saxony-Anhalt (Sachsen-Anhalt) in the west.

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

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

The article is divided into the following sections:

Physical and human geography 40
The land 40
Relief
Drainage and soils
Climate
Plant and animal life
Settlement patterns
The people 49
Ethnic divisions
Linguistic diversity
Religious profiles
Demographic trends
Population distribution
The economy 51
Administration of the economy
The former East German system

Resources
Agriculture, forestry, and fishing
Industry
Finance
Trade
Transportation
Administration and social conditions 58
Government
The former East German system
Justice
Political institutions
Armed forces
Law enforcement
Education
Health and welfare
Housing
Standards of living
Cultural life 64
The cultural milieu
State of the arts
Cultural institutions
History 67
Ancient history 67
Coexistence with Rome to AD 350
The migration period
Merovingians and Carolingians 69
Merovingian Germany
The rise of the Carolingians and Boniface
Charlemagne
The emergence of Germany
Germany from 911 to 1250 70
The 10th and 11th centuries
Germany and the Hohenstaufen, 1125–1250
Germany from 1250 to 1493 78
1250 to 1378
1378 to 1493
Germany from 1493 to *c.* 1760 87
Reform and Reformation: 1493–1555
The confessional age: 1555–1648
Territorial states in the age of absolutism
Germany from *c.* 1760 to 1871 98
Germany to 1815
The age of Metternich and the era of unification: 1815–71
Germany from 1871 to 1945 111
The German Empire, 1871–1918
The German Republic, 1918–33
The Third Reich, 1933–45
The era of partition 123
The Federal Republic of Germany
The German Democratic Republic
Germany reunited 129
Bibliography 129

PHYSICAL AND HUMAN GEOGRAPHY

The land

In its major topographical features, Germany shares with the countries on its eastern and western borders a profile that ascends in altitude from north to south. The North German Plain, or Lowland, constituting approximately the northern third of the republic, is but a segment of a vast lowland coastal area extending from extreme southwestern France across the Low Countries and throughout the Baltic areas to the Soviet Union. The central third of the territory, commonly known as the Central German Uplands, is a complex system of forested lowland mountain ranges, river basins, and plateaus. These follow a sequence of upland regions that range in a wide arc from the Massif Central of south central France, through Belgium and Luxembourg, across Germany, and through the entirety of Czechoslovakia to the Carpathian Mountains of eastern Europe. The southern third of Germany is, in effect, a northeastern extension of the Jura mountain system, which has its origins on the eastern bank of the Rhône valley in southeastern France. Entering Germany at the juncture of the French and Swiss borders in the extreme southwest, the Jura branches into two ranges within the country. South of the Jura Mountains and the basin of the Danube River, an undulating upland plateau forms the southernmost region of Germany as it skirts the country's southern frontiers and terminates at the shores of Lake Constance (Bodensee), the mountainous Allgäu, and the precipitous rise of the Bavarian Alps south of Munich.

The upland regions

RELIEF

The North German Plain, or Lowland. Embracing all of the *Länder* of Schleswig-Holstein, Bremen, Hamburg, and Mecklenburg-West Pomerania, as well as Berlin, the greater portion of Lower Saxony, Saxony-Anhalt, and Brandenburg, and the northwestern areas of North Rhine-Westphalia, the North German Plain presents, in spite of some intimations of monotony, an area of considerable variety in topography—to say nothing of scenic beauty. Although great stretches of the coastal plain between the Weser River and the Dutch border are largely flat, the moraine-formed areas stretching southward down the Jutland Peninsula of Denmark include gentle, rolling terrain of low hills and broad valleys, woodland areas, lakes, marshland, and heaths. Of the last, the best-known is the Lüneburg Heath, a plateau extending on a morainic belt between Hamburg and Hannover.

The rocks underlying the plains are Pleistocene sediments (less than 1.6 million years old). Underlying older rocks include Tertiary deposits, with lignite (brown coal) mined in the east, and Mesozoic sedimentary rocks, which lie too deep for easy exploitation. Exceptions include limestone quarries east of Berlin and ancient rock-salt and potash deposits, which can be mined because of upthrusts associated with salt-dome formation.

The surface of the plains was shaped largely by Ice Age forces. A northern morainic belt extends along the Baltic Sea, with ridges in the area north of Berlin, and the Fläming Hills, to the south, rising to some 330 feet (100 metres) above the general level. The coastline of Schleswig-Holstein and Lower Saxony along the North Sea is characterized by sandy dunes, tidal inlets, and marshland. Farther inland, the grassy dairylands are interspersed with the sudden rises of geest, oblong hills representing groups of old alluvial deposits. North of a line from Guben, on the Polish frontier, to Brandenburg, west of Berlin, the land was molded by southward-moving glaciers, which were formed during the Weichsel (or Vistula) glacial stage (the last Ice Age advance). The area south of this line was affected only by the earlier Saale and Elster glacial stages. In the northeastern region, terminal moraines mark the retreat phases of the ice sheet. South of the moraines are plains of outwash sands, and within them are sheets of till (ground moraine) littered with blocks and studded with innumerable lakes and ponds. To the southeast the older drift is generally leveled; there the leached soils, like those of the outwash sands and terminal moraines farther north, are generally poor and often forested, although the till plains south of the Baltic Sea offer good if heavy soils. The southern edge of the lowlands is marked by a belt of mainly windblown loess, dating from the Weichsel glacial stage, that extends from the Harz foreland to Upper Lusatia (Oberlausitz), with a tongue entering the Thuringian Basin. Here are found some of the country's best soils.

Off the western peninsula, the Schleswig portion of the Jutland coast, lie the North Frisian Islands, popular as summer resorts; the best-known of them is the isle of Sylt. Extending east to west off the coast of Lower Saxony lies the chain of the East Frisian Islands; their West Frisian extension belongs to The Netherlands. The tiny rocky island of Helgoland is located in the North Sea some 40 miles (65 kilometres) northwest of the mouth of the Elbe at Cuxhaven. The alluvial plain stretching from the eastern part of Holstein and the northern portion of North Rhine-Westphalia to the Dutch border is of great economic importance because of its rich soil and mineral wealth. Lignite and bituminous coal are found in areas west of the Elbe and in the lower Rhine areas; iron ore,


Scale 1:3,920,000
1 inch equals approx. 62 miles
0 25 50 75 mi
0 20 40 60 80 100 120 km
Conic Projection
7°
9°
11°
13°
15°
55°
53°
51°
49°
DENMARK
NORTH SEA
BALTIC SEA
NORTH FRISIAN ISLANDS
SYLT
HELGOLAND
FEHMARN
RÜGEN
Mecklenburg Bay
SCHLESWIG-HOLSTEIN NATIONAL PARK
NIEDERSACHSISCHES WATTENMEER NATIONAL PARK
EAST FRISIAN ISLANDS
Wadden Sea
Dollart Bay
Jade Bay
Elbe Bridges
Kiel Canal
Elbe
Weser
Ems
Oder
Oder-Spree Canal
Hohensaaten-Friedrichsthal Canal
Elbe-Havel Canal
Mittelland Canal
Lake Dümmer
Lake Steinhude
Lake Müritz
Lake Kummerow
Lake Schwerin
Lake Plau
Helpter Hill 179
LÜNEBURG HEATH
MECKLENBURG
PRIGNITZ
UCKERMARK
ALTMARK
FLÄMING HILLS
TEUTOBURG FOREST
MÜNSTERLAND
SAUERLAND
ROTHAAR HILLS
SEVEN HILLS
WESTER FOREST
EIFEL UPLAND
HUNSRÜCK UPLAND
TAUNUS MTS.
VOGELS MOUNTAIN
RHÖN MTS.
WASSER MT.
SPESSART
HARZ MTS.
KYFFHÄUSER MTS.
THURINGIA
Mount Großer Beer 982
FRANCONIAN FOREST
FICHTEL HILLS
Mount Schnee 1051
ORE MOUNTAINS
Fichtel Mountain 1214
VOGTLAND
LAUSITZ
OBER LAUSITZ
FRANCONIAN JURA
SWABIAN JURA
BAVARIAN FOREST
BAVARIAN FOREST NATIONAL PARK
BOHEMIAN FOREST
BLACK FOREST
KAISERSTUHL MASSIF
Mount Feldberg 1493
HAARDT MTS.
PALATINATE FOREST
BAVARIAN ALPS
BERCHTESGADEN NATIONAL PARK
Zugspitze 2962
Helpter Hill 179
Hamburg
Bremen
Hannover
Berlin
Dresden
Leipzig
Magdeburg
Dortmund
Essen
Duisburg
Düsseldorf
Cologne
Bonn
Frankfurt am Main
Wiesbaden
Mannheim
Karlsruhe
Stuttgart
Nürnberg
Munich
Augsburg
Regensburg
Würzburg
Darmstadt
Kassel
Erfurt
Chemnitz
Zwickau
Prague
CZECHOSLOVAKIA
POLAND
AUSTRIA
SWITZERLAND
LIECHTENSTEIN
FRANCE
BELGIUM
LUXEMBOURG
THE NETHERLANDS
Lake Constance
Danube
Rhine
Elbe
Oder
Weser
Ems
Main
Neckar
Saar
Moselle
Inn
Isar
Lech
Naab
Regen
Regnitz
Warnow
Recknitz
Peene
Spree
Schwarze Elster
Elster
Saale
Mulde
Neisse
Lahn
Dill
Fulda
Werra
Aller
Leine
Havel
Ruhr
Lippe
Nahe
Mosel
Ill
Danube
Lake Ammer
Lake Starnberg
Lake Chiem
Lake Tegern
Königssee
©Encyclopædia Britannica Inc.
Cities over 1,000,000
Cities 500,000 to 1,000,000
Cities 100,000 to 500,000
Cities under 100,000
National capitals
Political subdivision capitals
Political subdivision names
International boundaries
Political subdivision boundaries
Canals
Dams
Swamps and marshes
National parks
Spot elevations in metres (1 m = 3.28 ft)

MAP INDEX

Political subdivisions

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

Cities and towns

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

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

Physical features and points of interest

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

potash, and salt deposits are found in the plains north of the Harz Mountains, and, in addition, there are some oil and natural gas deposits between the Ems, Weser, and Elbe rivers and up through Schleswig-Holstein.

Apart from its agricultural importance, particularly in grain, dairy products, and livestock, the plain's wealth of rivers and canals—largely a result of its glacial origin—is an important adjunct to Germany's intricate inland-waterway transportation system. The Elbe and Weser rivers flow through the country's two great inland seaports, Hamburg and Bremen, respectively. The Ems, running parallel to the Dutch border, also empties into the North Sea. Farther to the south and moving to the northwest, the Rhine River, Europe's greatest inland waterway, flows to the North Sea through The Netherlands after passing Cologne and Düsseldorf, just to the west of Europe's largest concentration of industrial and mineral wealth, the Ruhr. This district, taking its name from a river valley that constitutes only a portion of the area, represents a great industry-population complex roughly triangular in shape. At its southern tip lies Cologne, with Dortmund in the northeast and Duisburg in the northwest. Rich in coal and formerly in iron, it is the former western sector's greatest centre of heavy industry—steel, machinery, oil refining, chemicals, glass, ceramics, textiles, and many other manufactures.

Industrial concentration in the Ruhr

The Central German Uplands. In general, the Central German Uplands, a portion of the Central European Uplands, constitute an extremely intricate and geologically complex continuum of mountain ranges, separated in their larger divisions by river and stream valleys. They are rounded and of no great height, most less than 3,500 feet (1,100 metres) in elevation. At their upper levels they tend to be heavily forested; they are important sources of timber, and a number of these mountain districts are known as forestlands rather than mountain ranges. A number of areas, especially those formed by volcanic activity in relatively recent geologic times, are important sources of minerals and centres of local industry. Valleys are utilized for crop raising and, where climatic conditions permit, for growing wine grapes. In other areas, a plentiful

supply of hydroelectricity and easy access to water, rail, and road transport support mining, manufacturing, and processing industries.

The core of the whole region is hard, crystalline rock associated with the ancient Hercynian mountain-building period (some 300 million years ago), when molten granitic intrusions formed ore deposits. Toward the end of the Paleozoic era (about 245 million years ago), this eroded mountain range was filling its associated structural depressions with debris, including some hard-coal seams that are no longer worked. In the ensuing Mesozoic era, layers of limestone, sands, and clays, along with deposits of potash and salt, covered the older rocks. Tertiary earth movements shattered this ancient massif, and some blocks were uplifted to form mountains. The Mesozoic cover was stripped from the mountains by erosion, surviving only in the basins, such as the Thuringian Basin and the Harz foreland. Basin soils, mostly brown earths, vary according to the outcrop concerned; mountain soils are generally poor, thin, or peaty.

This Central German Upland district—all or most of Hesse, Rhineland-Palatinate, the Saarland, Thuringia, and Saxony and the southern portions of North Rhine-Westphalia, Lower Saxony, Saxony-Anhalt, and Brandenburg—is extremely variegated in character, sharing only its mountainous and hilly terrain as a common feature. It contains some of Germany's most centrally located and wealthy industrial areas (in the Rhine-Main area of southern Hesse) and also some of the most remote and sparsely populated backwoods regions.

The high tier. The high tier of the Central German Uplands extends east of the Rhine and south of the Ruhr and the Lippe rivers and north of the Lahn River eastward to the upper reaches of the Weser. In this range, or bordering on it, are the Teutoburg Forest in the north, the Seven Hills and Sauerland in the west, and the rugged Westerwald in its south near the Lahn. Across the Weser and its tributary, the Werra, the range continues into the Harz region, south of Hannover and Braunschweig. The Harz Mountains are noted for their beauty and popular as a resort area. The Brocken, at 3,747 feet (1,142 metres) the highest peak in the Harz range, is known for its many associations with German legend; its peak lies just to the east of the former inter-German border. Also in the former eastern sector are the Ore Mountains (Erzgebirge), the mountains of Upper Lusatia and the Thuringian Forest, the rolling hills of the Thuringian and Erzgebirge basins, and, near Dresden, the lowlands of the Elbe valley.

The Harz Mountains

East of the Rhine and between the Lahn and Main rivers are the highland districts of the Taunus, also a favoured resort. Between the Main and Kinzing rivers lies the Spessart, an extensive forest preserve in which the last of Germany's ancient oak trees are found. To its southwest, between the Main and Neckar rivers, lies the fabled Oden Forest (Odenwald), a remote region near the populous Frankfurt am Main metropolitan area; in this wood the legendary Siegfried is said to have been slain.

To the northwest, west of the Rhine and north of the Moselle (German: Mosel) River, the vast tract of the Eifel Upland continues into Belgium as the Ardennes. This range, together with the parallel mountain ridge south of the Moselle known as the Hunsrück, bears evidence of geologically recent volcanic action and is an important source of basalt. Semiprecious stones are mined in the vicinity of Idar-Oberstein. South of the Hunsrück lie the highlands of the Pfälzer Mountains. The Saar basin, bordering French Lorraine in the extreme southwest of the Pfälzer Mountains, is, like the Ruhr, rich in coal and iron, though of an even higher quality than that of the northern mining regions.

The Rhine gorge. The Rhine gorge cuts and winds through the plateau of lowland mountains that represents a continuum between the Eifel and the Hunsrück to the west and the Seven Hills and Taunus to the east. This famed and most spectacular stretch of the Rhine, from about Andernach in the north to Bingen at its southern reach, is lined with hilltop castles, vineyards along its precipitous slopes, and a sequence of picturesque towns along its banks. The winding stretches near Sankt Goarshausen pass between all but vertical rocky cliffs, the most famous of which is the ominous Lorelei rock.

The Southern Highlands, or plateau region. The Main River, which, upstream from its mouth at the Rhine near Mainz, meanders in a westerly direction from near the westernmost extent of the Czechoslovakian border, may be regarded as the dividing line between the Central German Uplands and the mountainous and plateau region of southern Germany. The area lacks any specific characteristic to distinguish it from the Central German Uplands. In its topography, if not in its external appearance, it is rather similar to the central regions. Its mountains, however, are derived from different systems from those in central Germany, and its mean elevation above sea level, even in plateau regions, is considerably higher.

In an area of northern Baden-Württemberg and northeastern Bavaria framed by the basins of the Main and Neckar rivers and extending to the east is a region of low hills, plateaus, and valleys. It is both rich as an agricultural

E. Streichan—Shostal Associates/Superstock



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

area and noted for its many ancient towns that, largely because of their remoteness from the main routes of transport, have remained in their outward appearance relatively untouched by modern development. This district includes the renowned "Romantic Road" from Donauwörth to Würzburg, passing through Rothenburg ob der Tauber, a carefully preserved walled city that has remained virtually unchanged since the 17th century.

The Jura range and the Black Forest

The northernmost reaches of the Jura range, entering Germany from Switzerland in the extreme southwest, branch into two forks. The western fork, moving in a northerly direction parallel with the Rhine River to a point just south of Karlsruhe, is commonly known as the Black Forest (Schwarzwald) because of the dark appearance of the area's giant fir trees. The Black Forest is rich in thermal water sources and abounds in resort areas, spas, and sanitoriums. West of the Black Forest is the wide, sloping eastern valley of the Rhine opposite Alsace, in which the lower and smoother Vosges Mountains on the French side run parallel to the Black Forest as part of the northern Jura.

The eastern fork of the Jura range branches off from the northern extension some 20 miles east of Freiburg im Breisgau, where the Danube River rises to meander 1,770 miles eastward past seven countries to the Black Sea. This branch extends due east some 320 miles to near the source of the Main, skirting to the south of Stuttgart and Nürnberg and following the northern basin of the upper Danube as far as Donauwörth. Its southern portion is known as the Swabian Jura (Schwäbische Alb) and its northern portion as the Franconian Jura (Fränkische Alb); some parts of it evidence curious and often rugged geologic formations.

From Donauwörth, the Jura swings in an arc almost due north to its terminal point near Bayreuth. To the east, toward the Czechoslovakian border, lie the rolling wooded plains of the Bohemian Forest; farther south, near Regensburg, the Bavarian Forest lines the northern slopes of the Danube valley.

Justly famed as they are for their dramatic beauty—and especially for the foothills in their immediate foreground—the Bavarian Alps represent, in effect, only the northern edge of two relatively minor extensions of the Austrian Alps, the Arlberg and Tirolean ranges. Germany's share in Alpine territory is, in fact, miniscule by comparison with that of Austria, Switzerland, Italy, and France.

DRAINAGE AND SOILS

Drainage. Because of Germany's steady upward incline from the coastal regions of the north to the Bavarian Plateau and Alpine regions of the southeast, the greater majority of its rivers and their tributaries flow north and empty into the North Sea via the Rhine, the Ems, the Weser, and the Elbe. The northward-flowing Oder (with its tributary the Neisse) drains the extreme eastern frontier before entering the Baltic Sea through Poland. While many of the tributaries to these major rivers meander east-west or west-east, only the Lahn flows in a southerly direction, emptying into the Rhine. The remainder of German waters—all of which are located in Bavaria—escape into the Danube River on its eastward course toward the Balkans; the main tributaries to the south of the Danube are the Iller, the Lech, the Isar, and the Inn; flowing southward from the north are the Altmühl, the Naab, and the Regen.

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

Germany has relatively few lakes. The greatest concentration comprises the shallow lakes of the postglacial lowland of the northeast, including the Berlin area. The largest natural lake of the region is Lake Müritz (44 square miles) in the Weichsel glacial drift of Mecklenburg-West Pomerania. In addition to Lakes Dümmer and Steinhude in Lower Saxony, a few small lakes of glacial origin dot Schleswig-Holstein. The remainder of Germany's lakes are concentrated at the extreme southeastern corner of Upper Bavaria, many of these in outstandingly beautiful surroundings. Germany shares Lake Constance, its largest lake (having the proportions of an inland sea), with Switzerland and Austria. Stretching 40 miles at its longest reach, Lake Constance is, in effect, an elongated basin filled by the Rhine River from its point of origin in eastern Switzerland; at its southwesternmost fork the lake reassumes the form of a river and flows eastward toward Basel.

Lake Constance

Soils. Although most of Germany's arable land is under cultivation, comparatively few areas are highly fertile. The terrains thus unsuited for large-scale grain production are given over to grazing land. In the lowlands of the north, the land overwhelmingly comprises sand and loam, though along the North Sea littoral up the western coast of Jutland the soil is predominantly sandy. The eastern half of the Jutland Peninsula and the tier of the Baltic coast extending southward to Hamburg is mostly the clay of the northern morainic belt, or Baltic Ridge.

An especially fertile area known as the Börde is covered with a siltlike loess; it extends in a wide belt from west of the Rhine near the Ruhr district eastward, following the southern edge of the Mittelland Canal and extending across the Elbe near Magdeburg. Other regions containing this rich soil are found in scattered areas, particularly in the lowland regions west of the Rhine to the Dutch border and in three districts around the greater Frankfurt am Main area. Other loess-covered districts of considerable fertility are located in the north of Württemberg, in small portions of northern Bavaria, and in wide ranges to the northeast of Ulm and Munich. Some marshland is found both in the North German Plain and in the Alpine plateau.

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

CLIMATE

Germany is favoured with a generally temperate climate, especially in view of its northerly latitudes and the distance of the larger portions of its territory from the warming influence of the Gulf Stream in the Atlantic Ocean. Extremely high temperatures in the summer and deep, prolonged frost in the winter are rare. These conditions, together with a more than abundant and well-distributed amount of rainfall, afford ideal conditions for crop raising. As throughout western Europe in general, however, Germany's climate is subject to quick variations when the warm westerly maritime climate emanating from the Gulf Stream collides with the more extreme climatic ranges moving in from northeastern Europe. The exposed North and Baltic sea coastlines enhance the maritime component.

Variations in seasons

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

The intricate topography of the Central German Uplands and Southern Highlands and their valleys causes infinite local variations in temperature, frequency of sunshine, humidity, and precipitation. For this reason Germany produces more vintages and varieties of wine than any other country, at least 10,000 registered for marketing purposes.

The northwestern and lowland portions of the republic are affected chiefly by the uniformly moist air, moderate in temperature, that is carried inland from the North Sea by the prevailing westerly winds. While this influence, on the whole, affords moderately warm summers and mild winters, it is accompanied by the disadvantages of high humidities, extended stretches of rainfall, and, in the cooler seasons, fog. Precipitation diminishes eastward as the plains open toward the Eurasian interior, and the average temperatures for the warmest and coldest months become more extreme. The hilly districts of the central and southwestern regions and, to an even greater degree, the upland and plateau districts of the southeast are subject to the more pronounced ranges of hot and cold from the countervailing continental climate. The mountains have a wetter and cooler climate, with westward-facing slopes receiving the highest rainfall from maritime air masses. At a station on the Brocken, annual precipitation reaches 63 inches at an altitude of 3,747 feet (1,142 metres). The sheltered lee slopes and basins have, by contrast, rainfall that is extremely low (Alsleben, east of the Harz Mountains, receives 17 inches annually) and hot summers (July mean temperatures above 64° F [18° C]), necessitating crop irrigation. Southeastern Germany may intermittently be the coldest area of the country in the winter; but the valleys of the Rhine, Main, Neckar, and Moselle rivers may also be the hottest in the summer. Winters in the North German Plain tend to be consistently colder, if only by a few degrees, than in the south, largely because of winds from Scandinavia.

Mountain climate

One anomaly of the climate of Upper Bavaria in the warm seasons is the occasional appearance of warm, dry air passing over the northern Alps to the Bavarian Plateau. These mild winds, known as foehns, not only create an optical phenomenon that makes the Alps visible from points where normally they would be out of sight but also are reputed to have a depressing psychological effect upon the inhabitants.

Annual mean precipitation varies according to region. In the North German Plain it ranges from 20 to 28 inches (510 to 710 millimetres); in the Central Uplands, from 28 to 59 inches; and in the Alpine districts, up to and exceeding 78 inches. The average temperature for January, the coldest month, varies from 27° to 34° F (−3° to 1° C) in the lowland areas of the north; in the upland regions, depending on altitude, temperatures in January fall to an average of 21° F (−6° C). July temperatures average from 60° to 66° F (16° to 19° C) and run slightly higher in the sheltered river valleys.

(G.H.K./H.-J.K./R.E.H.M./T.H.El./Ed.)

PLANT AND ANIMAL LIFE

A somewhat arbitrary south-north slice across central Europe, Germany does not have vegetation and animal life greatly distinct from that of surrounding countries. In its natural state, the region would be almost entirely forested. The glaciated lowlands were initially covered by a predominantly deciduous mixed forest of oak, linden, beech, and birch. The richest habitats were provided by the oak-hornbeam woodlands of the loess belts, their canopies sheltering numerous insects, birds, and the red squirrel. The rich shrub layer attracted red deer, roe deer, and wild pigs, among the larger mammals, as well as badgers, foxes, and hamsters and numerous other rodents.

Plants. Only beech and birch remain evident in some abundance in the patchy forests of the North German Plain, Schleswig-Holstein, and the rim of the Thuringian Basin; oak is found less frequently. After the felling of these primeval forests, Germany took the lead in systematic reforestation, mostly, however, by planting the faster-growing conifers. Thus, fir, spruce, and pine outnumber deciduous trees of all varieties by more than two to one. Because of a carefully regulated system of reforestation, Germany contains the greatest density of woodland of any larger region of western Europe exclusive of Scandinavia, and forestry and timber production are major facets of German agriculture. The heavily forested districts, nevertheless, are almost solely confined to the central and southern uplands. Industrial pollution, particularly after decades of unregulated output in the Democratic Republic and neighbouring countries of the former socialist bloc, has severely compromised Germany's forest resources.

Coniferous forests

Fruit trees—apple, pear, peach, cherry, plum, apricot, and quince—and the walnut, chestnut, and hazelnut trees are widely distributed, but, like the vineyards, they thrive best in the western and southwestern regions. Strawberries, blueberries, blackberries, cranberries, and currants abound in the forests and meadows of all districts.

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

Although the white stork, the cormorant, the horned owl, the fishhawk, and the osprey have nearly become extinct, the golden eagle has gained in numbers in the Alpine districts, and the heron still holds forth in scattered areas of the country. The bustard and eagle owl are still found along the Elbe.

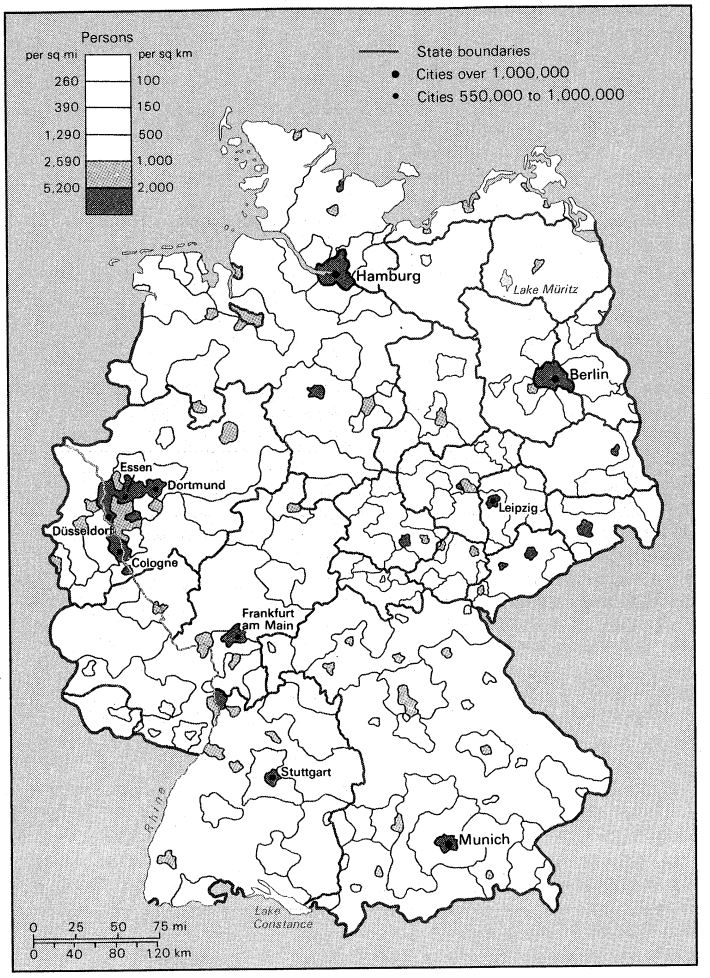
Parks and preserves. The republic has established national parks in the Bavarian Forest and at Berchtesgaden and elsewhere. Several major resort areas, such as portions of the Black Forest, have been given the status of natural park. In addition, the government has declared several thousand localities as nature reserves or as protected scenic areas. Thousands more locations of unusual historical interest or natural beauty have been set aside as natural monuments.

SETTLEMENT PATTERNS

The boundaries of the modern *Länder* of Germany were established in comparatively modern times. The more than 2,000 political units, large and small, that existed under the Holy Roman Empire gradually were consolidated through war, treaties, dynastic marriages, large-scale amalgamations effected under Napoleonic rule, further consolidations brought about by the establishment of the German Empire, or Reich, in 1871, and, in the wake of World War II, by plebiscite. Transcending and often defying these modern political borders, however, are the traditional regions of Germany, the characters of which are chiefly ethnic in nature and often prehistoric in origin. These regions have persisted with remarkable tenacity throughout the turbulent history of the German territories and the radical social upheavals of the 20th century. In addition to what might be described as the "tribal" cohesion of the regions, the affinities of linguistic dialect, traditional political allegiances, and, often as not, religion will in different measure serve not only to set off one region from another but also to account for the various subdivisions within the major regions themselves.

Regional feelings are still very strong in Germany. Just as a Briton of Celtic stock may call himself a Scotsman or Welshman, even though his family has resided for years in England, or as an eighth-generation American will often describe himself as Scotch-Irish or Dutch, a German will claim descent from the region of his ancestors no matter how remote the ties may be. To this persistent consciousness of ancestral origins must be affixed the German's deeply felt sense of *Heimat,* his "homeland" (or often the homeland of his forebears), which connotes a

The sense of *Heimat*



Population density of Germany.

much narrower sense of native region, namely, the valley or forestland, plain, mountain, or even village or town where one's ancestral ties run deep. These feelings were dissipated, however, by the enormous influx after World War II of Germans from East Germany and the former provinces now incorporated into Poland.

Many of the regional names are no longer reflected in any formal political designation, but they are still used in everyday parlance. In many instances, political consolidations, especially in the south, have not only cut across the conventional boundaries of the ancient regions but also have brought together under a single political administration regions that were, at best, disparate and, at worst, disdainful or even hostile to one another.

Southern regions. In the south the major regions are largely within the *Länder* of Baden-Württemberg, Bavaria, Saarland, and Rhineland-Palatinate. Baden-Württemberg extends from an area around Mannheim southward along the Rhine, through the Black Forest to the Swiss border, and east to Ulm. The natives of Baden-Württemberg have close ties and affinities with the northwest area of German-speaking Switzerland and with Alsace (France), the former German province across the Rhine; like the Swabians, they are of ancient Alemannic stock. Northwest of the Baden-Württemberg region, stretching north along the Rhine is the Rhineland-Palatinate *Land,* traditionally known as the Palatinate (Pfalz). The Saarland is a region unto itself, but it has characteristic aspects of both the surrounding Rhineland-Palatinate and the neighbouring French region of Lorraine, which used to be a province of Germany.

Swabia (Schwaben) is a notable example of a former region whose extent transcends *Länder* boundaries. It embraced most of the former states of Württemberg and Hohenzollern, but its eastern portion between the Iller and Lech rivers, roughly between the cities of Ulm and Augsburg, was ceded to Bavaria in 1806. The Schwaben *Regierungsbezirk* (administrative district), in a pocket in the extreme southwestern corner of Bavaria, not only embraces part of the traditional Swabian region but also has close historic ties with the Vorarlberg state of Austria, itself settled by Alemannic tribes.

The traditional regions of Bavaria and Franconia

The region popularly thought of as Bavaria—famed for its traditional Alpine dress, the dirndl for women and lederhosen for men, and renowned for gargantuan consumption of beer to the accompaniment of music from a brass band—is more correctly Upper Bavaria. It comprises the southeastern quadrant of Bavaria, Germany's largest *Land*—east of the Lech and south of the Danube River eastward from Donauwörth. The northern and geographically larger portion of the *Land* is almost completely occupied by the traditional region of Franconia—its name deriving from its East Frankish settlers. Upper Franconia takes in the highland district along the Czechoslovakian and former inter-German borders near Bayreuth; Central Franconia is located in the areas surrounding Nürnberg, while Lower Franconia centres around Würzburg. Portions of northeast Baden-Württemberg *Land* are also Franconian in origin and character.

Northern regions. The regions north of the Main River, with certain exceptions, are larger and more homogeneous, except for the inroads of modern internal migration in such large urban districts as the Rhine-Main and Rhine-Ruhr metropolitan areas. Only the Rhineland, consisting of the areas of the west bank of the Rhine River from Mainz to Cologne, differs notably from the regions north of the Main. The Moselle and Eifel districts to its west are of a different cast. On the opposite bank of the Rhine, the state of Hesse, in spite of its relatively large size, is one of the few states that can claim an almost complete identity of *Land,* region, and ethnic origin; the Hessians are a branch of the Franks in the ancient land of the Chatti.

The Rhineland, by contrast, is politically wedded today in North Rhine-Westphalia to the very different regions of the Sauerland, east of the Rhine near Ebbe, and to Westphalia, with its close affinities in character and outward appearance to the eastern provinces of The Netherlands. The Ruhr itself is sometimes called the *Kohlenpott* ("coal pot"), a derisive reference to the modern migration of workers from the German eastern provinces and Poland to tend its mines and factories.

Lower Saxony *Land,* apart from incidental subregions demarcated by landscape, such as the Lüneburg Heath and the Harz, divides roughly into the traditional regions of Hannover and Oldenburg, except for its coastal area, East Friesland, which has close ties to West Friesland in The Netherlands and North Friesland off the coast of Jutland. The slip of coastline extending from the Dutch border up around the shore of the North Sea to Schleswig-Holstein is also sometimes referred to by the Low German term *Waterkant,* "water's edge." The ancient Hanseatic cities of Bremen, Hamburg, and Lübeck may be ranked as separate regions in their own right; all have in common a sense of self-sufficiency, independence, and aloofness from the remainder of Germany. The amalgamated *Land* created from the provinces of Schleswig and Holstein, once duchies under the Danish crown although not settled by tribes of Danish stock, retains obvious ties of kinship and culture that extend up the peninsula to its Jutlandic cousins across the border in Denmark.

East of the Elbe and Saale rivers, population pressure in the 12th and 13th centuries drove Germans to settle Slav lands, where regularly planned settlements predominated. The villages of the northern lowlands east of the Elbe shrank under the impact of wars and famine to communities with fewer than 100 inhabitants, dependent on the Junker landlords who swept up their lands into manorial farms (*Gut*). Old settlement patterns prevailed until World War II, when devastation and the advent of socialism remolded both the urban and rural landscapes.

Varieties of the people. Although such generalizations fail to account for the exceptions, it can be stated that the people of the regions north of the Main River tend more toward the popular notion of the blond, blue-eyed, sobersided, hard-working, and often dour German of legend.

Contrasts among peoples

A conspicuous exception to this stereotype would be that of the Rhinelander, noted for his carefree disposition and almost Gallic affinities with Latin Europe. The Franconians in the south also share in the north German profile of being rather sturdily given to work and no nonsense, and they are rather conservative in the preservation of their traditional ways. A great contrast is found between the easygoing natives of Baden and their fellow Alemannians the Swabians, who among other Germans are thought to be the hardest-working, the thriftiest, the most inclined to introspection, and one of the most self-sufficient of German peoples. No greater regional contrast could exist than that between the Swabians and the Bavarians, the latter noted for their garrulous and sometimes coarse temperament, rough humour, and vigorous pursuit of fleshly pleasures. Nonetheless, if Swabia has furnished Germany with its poets and great intellects, Bavaria is the home of its artists.

Rural Germany. The manner in which land under cultivation in Germany is divided and the major patterns that have prevailed in the utilization of crop and grazing lands and timberland have been determined not only by the somewhat severe limitations of the soil but also by regional convention and historical circumstance and, especially in modern times, by economic expediency. The disposition of arable land and the structure of rural settlement underwent steady change over the centuries, as circumstances dictated, until the major land reforms and the onset of industrialization in the 19th century. The most radical and abrupt adjustments in agriculture and the rural population occurred after World War II.

Historic methods of farming and land use

Under the traditional patterns that prevailed throughout medieval and early modern times, to some degree even until the mid-20th century, the simplest structural form of farmland was that of the *Einödflur,* or *Einzelhof,* the individual farm unattached to a community or village. This type of farming in detachment has been rather the exception in Germany, and it is widespread chiefly in the grazing lands of the northwest. Numerous variations are encountered in the more complex patterns, usually centred around a village, often in conjunction with a large estate. To these variations must be added the nature of tenancy—whether the land was the property of a *Gutsherr,* a large landholder, usually of the nobility; a *Grossbauer,* a peasant or farmer with large holdings; a *Kleinbauer,* a small farmer-proprietor; or a *Pächter,* a tenant farmer, large or small.

The ancient division of land, the *Hufe,* comprising 60 to 100 acres (25 to 40 hectares), was conceived of as the amount of land, arranged into one or several strips, that was necessary for growing enough crops or for maintaining sufficient livestock to sustain one farming operation. In the larger village enterprises and especially common in most of the regions of former West Germany, the *Hufen* were arranged into *Gewannfluren,* or the acreage surrounding a village to which the farmers went from their villages to tend their crops on the strips of land allotted to them or, in some instances, on the communal property of the village. Of this latter form, large tracts of forestland are still under the ownership of the *Gemeinde* ("community").

The patterns of the *Hufenfluren* ("acreage") differed widely, depending on the type of soil, the terrain, and the region. As often as not, a farmer's land would consist of several plots or strips frequently separated from one another. The divisions could also be quite irregular and uneconomically small. For this reason, a major reform in West German agriculture in the post-World War II era was the attempt to bring about a more efficient arrangement in the division of agricultural lands through a series of consolidations and redistribution of acreage. Along with this realignment of agricultural properties, largely accomplished by land sales and exchanges, came the gradual disappearance of the small farmer and of multicrop or self-sustaining farms in favour of the large-scale farmer concentrating on one or a few selected crops, the nature of which is likely to be determined by planning requirements at the *Land* or *Bund* (federal) level and adjusted to fit the crop quotas and price structures indicated by the Federal Republic's membership in the European Communities. As in other industrial countries, the migration from the land continues.

Village patterns

German rural settlement follows certain set patterns—from the relatively infrequent individual farm to the *Weiler,* a loose collection of farms found most frequently in the southwest and the Alpine forelands, to the many variations of the *Dorf* ("village"). By far the most common form of village in what was formerly West Germany, representing as it does the historically older portion of the German regions, is the *Haufendorf,* a conglomerate village of irregular pattern. Typically, it comprises a main thoroughfare and a square or village centre with such customary fixtures as a church, local administrative offices, marketing facilities, inns, shops, and the like and an asymmetrical patchwork of streets and roads leading from the centre.

Other patterns of village structure include the *Rundling,* a radial village (the form of which was determined by economic function and not, as was commonly supposed, by considerations of defense), and the *Reihendorf,* a linear village, the simplest form of which is the *Strassendorf,* in which the houses, buildings, and farm structures simply line the highway. These latter two forms are much more common in the former Democratic Republic, the lands of which were settled relatively later than those in western and southern Germany. Among the minor variants found in all areas of the west and south and determined by terrain are villages formed in forest clearings, mountainous areas, drained marshland, or moors. The transition of a village to the rank of a town historically was betokened by its being granted the privilege of the *Marktfleckenrecht,* the privilege of holding a market. Today, however, a village may lose its essentially agricultural character by the sudden appearance of a factory or an urban housing tract on its outskirts, or, especially in the southern regions, through a reorientation of its economy toward tourism.

The introduction of state and cooperative farms in the Democratic Republic beginning in 1949 radically changed the pattern of rural settlement. New buildings marked the introduction of mechanized cultivation or the large-scale farming of animals; multistory apartments and community centres reflected a new concept of rural life.

Urban Germany. After The Netherlands, Belgium, and the United Kingdom, Germany has the highest density of population in Europe. One-third of the population live in cities of 100,000 inhabitants or more, and two-fifths reside in places of fewer than 20,000 population, including villages or communities of 2,000 persons or fewer. Germany is essentially a country of small- and medium-sized towns and cities; only Berlin, Hamburg, and Munich exceed 1,000,000 people. Despite the heavy density of population, distribution is somewhat uneven. In addition, owing to divergent government policies during the partition, the former eastern and western sectors developed rather distinctive urban configurations.

About half of all residents of former West Germany live in or near one of the 10 major metropolitan centres, the most populous of which is the Rhine-Ruhr district—a conurbation including Düsseldorf, Essen, Dortmund, Duisburg, and Cologne and dubbed "Ruhr City." Apart from the "city-states" of Berlin, Hamburg, and Bremen, the other heavy concentrations of population are centred around Hannover, Frankfurt am Main, Mannheim-Heidelberg, Stuttgart, Nürnberg, and Munich.

Upon its formation in 1949, the German Democratic Republic inherited an extremely uneven urban distribution, with a marked southern concentration of industrial towns and cities, notably the conurbations of Halle-Leipzig, Chemnitz- (Karl-Marx-Stadt-) Zwickau, and Dresden. In the north, by contrast, the urban system took the form of widely spaced, small- or medium-sized market and administrative centres. The exception was Berlin, which by 1943 had reached its peak city population of nearly 4.5 million. The policy of the 1950s envisaged a move toward greater social-spatial equilibrium by the creation of new industrial sites in the north with associated new or expanded towns, as at Eisenhüttenstadt (Stalinstadt). This policy was not pursued, although there was considerable urban expansion on the Baltic coast, notably at Rostock,

while the northern *Bezirk* capitals expanded greatly. A characteristic of urban development, particularly in the 1970s and '80s, was the construction of peripheral high-rise housing projects for 50,000–150,000 people (as at Berlin-Marzahn, Halle-Neustadt, Leipzig-Gronau). East Berlin, the largest city, with its "showcase" reconstructed centre, grew steadily with the influx of predominantly young people. The impact of the events of 1989 and '90 on urban migration throughout Germany would be difficult to measure for some time.

The people

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

Interacting factors in cultural patterns

The areas of central Europe called Germania by the Romans (for only one of numerous tribes in the regions north of the Alps) were known as Germany (Deutschland) only in an imprecise geographic sense until the founding of the Reich in 1871. Until that time, Germany had only the loosest bonds of a multidialectal language, certain cultural affinities, disjointed political affiliations, and, to an even lesser degree, ties of kinship upon which to form any sense of nationhood or national consciousness. Certainly Germany was far short of being a nation-state in the manner of France or England. The ethnic and linguistic subdivisions of today's Germany of necessity transcend political boundaries, even though the Federal Republic, by and large, embraces most of the territories that made up the nucleus of the German nation—of both the Roman Germania and the Carolingian East Frankish kingdom.

The Germanic, or Teutonic, element represents only a portion of the ethnic profile of what are known as the German peoples. Account must be taken also of a considerable substratum of the pre-German Celtic inhabitants in the west and of numerous Slavic strains in the east. Whatever internal cohesion existed in the past within the borders of present-day Germany was obscured after World War II to some extent by the absorption into the western sector of several million expellees and refugees from the territories to the east. The Federal Republic has also acquired a substantial non-German minority of a few million *Gastarbeiter* ("guest workers"), migrant labourers most of whom are from the nations of southern Europe and Turkey. Permanent populations of foreign workers were officially prohibited in East Germany; nevertheless, some 90,000 workers from Vietnam, Angola, Cuba, and Mozambique were imported, ostensibly for "education and training." Polish day workers also moved across the eastern boundary.

ETHNIC DIVISIONS

The traditional ethnic divisions of Germany, coinciding in their larger dimensions with the major regions, go back to tribes and branches of tribes that had settled mainly there by the time of the Roman occupation. Their ultimate origin is still a matter of speculation if not dispute, but it is commonly held that they moved into central Europe from the Baltic regions and either drove the resident Celtic tribes farther west or partly absorbed them. The Celtic strain is notably marked by one of the most enduring forms of evidence: the place-names of rivers, mountains, settlements, and other landmarks that have come into the German nomenclature most often from the Latin rendition of the original Celtic term. The names of the Rhine, Danube, and Neckar rivers, to cite only the best-known, are derived from the Celtic.

Six major German ethnic groups have persisted in their historic regions from Roman and early medieval times to the present, and their numerous subdivisions are of no small historical, linguistic, and cultural importance. The southwest of the present-day Federal Republic was settled by the Alemannic tribes, who occupy not only the German traditional regions of Baden and Swabia but also the French province of Alsace, the German regions of Switzerland, all of Liechtenstein, and the Vorarlberg state of western Austria. The southeast, roughly south of the Danube River and east of the Lech River, was settled by the Bavarians, who share ties of kinship and linguistic dialect with the greater portions of modern Austria.

The central districts of Germany were occupied by the Upper Franks—the Main, Rhine, Moselle, and Hessian Franks—as well as the Thuringians. The Lower Franks find their present-day descendants in portions of North Rhine-Westphalia *Land,* Flanders, and The Netherlands. The Thuringians, moving eastward, colonized the areas encompassed by what are now Thuringia, Saxony, and Saxony-Anhalt *Länder.*

The greater portion of the North German Plain, inhabited by the "original" Saxons, is preserved in the names of the latter two *Länder* mentioned above, as well as that of Lower Saxony. The descendants of these ancient Saxons, however, have few kinship ties with the people of Saxony and Saxony-Anhalt. They made up one of the several west Germanic tribes of the North Sea littoral, including the Angles, who invaded Britain in the 5th century—hence, "Anglo-Saxon."

The sixth major group, still intact if now small in number, is that of the Frisians on the islands off the coasts of Jutland and Lower Saxony. They are closely related to the West Frisians in the northern portions of The Netherlands.

The German ethnic groups of the east had their origin in later colonizations from the west, beginning in the 9th century; they are known generally as the Mecklenburgers, the Upper Saxons, the Brandenburgers, the Silesians, the Pomeranians, the Prussians, and the Baltic Germans. During these eastern colonizations, the indigenous Slavic and Baltic peoples were absorbed and Germanized in great measure. To the south, Austria (German: Österreich; Eastern Realm, as both the Latin and German versions of its name indicate) represents an eastward colonization by the Bavarians, of whom the Tiroleans form an ethnic subgroup.

Minority groups in the population

Chief among the historical minority groups in Germany were the Jews, present in western Germany since early medieval times; they numbered about 800,000 in all of Germany prior to the genocide of World War II. Small cultural pockets of the original Slavic inhabitants of the regions colonized east of the Elbe River still exist in the Lusatia region, north of Dresden, and their language, Sorbian (Wendish), still survives. The persecutions of the French Protestants in the late 17th century brought large numbers of Huguenots to the German lands, especially to the cities of the northwest and to Berlin. The provinces of Schleswig and Holstein, duchies of the Danish crown until the 1860s, were inhabited largely by peoples of west Germanic as opposed to north Germanic or Scandinavian stock; thus, only a small minority of Danes, estimated at 3 percent of the population, still live there.

Various other minorities, large and small, that may be noted in German territory in relatively modern times include the Polish workers who migrated to the Ruhr in the late 19th century; small groups of Russian émigrés who settled in Germany, especially in Berlin, after the October Revolution in Russia; the countless numbers of "displaced" or "stateless" persons from eastern Europe and the Balkans who were stranded in Germany at the end of World War II; and the minority of *Gastarbeiter,* their families, and descendants, many of whom may well become permanent.

LINGUISTIC DIVERSITY

The dialectal divisions of Germany, once of conspicuous significance for the ethnic and cultural distinctions they implied, persist in spite of such leveling and standardizing influences as mass education and communication and in spite of internal migration and the trend among the younger, better-trained, better-educated, and more mobile ranks of society to speak a standard, "accentless" German.

The repository of dialectal differences thus lies more with the rural populace of the regions and the long-time native inhabitants of the cities.

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

The three major dialectal divisions of Germany—as with the ethnic divisions—coincide almost identically with the major topographical regions of the North German Plain, the Central Uplands, and the southern Jura, Danube basin, and Alpine districts. In reverse order, these dialectal divisions are known as Upper German, Central German, and Low German.

The dialects of German

Of the Upper German dialects, the Alemannic branch in the southwest is subdivided into Swabian, Low Alemannic, and High Alemannic. The first, the most widespread and still-ascending form, is spoken to the west and south of Stuttgart and as far east as Augsburg. Low Alemannic is spoken in Baden-Württemberg and Alsace, while High Alemannic is the dialect of German-speaking Switzerland. The Bavarian dialect, with its many local variations, is spoken in the areas south of the Danube River and east of the Lech River and throughout all of Austria, except in the Vorarlberg state, which is Swabian in origin.

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

The dialect known as Low German, or *Plattdeutsch,* historically covered all regions occupied by the Saxons and spread across the whole of the North German Plain. Although largely displaced by standard German, it is still widely spoken, especially among elderly and rural inhabitants in the areas near the North Sea and the Baltic, and is used in some radio broadcasts, newspapers, and educational programs. Tiny pockets of Frisian, the German dialect most closely related to English, persist.

RELIGIOUS PROFILES

Origin and political impact of religious division

Germany, ever since the Protestant Reformation and the ensuing Thirty Years' War, has notably been divided into the sectarian lines of Roman Catholic and Protestant, though in ways that far transcend considerations of religious doctrine and practice. Following the precept formulated in the Peace of Augsburg (1555), whereby the inhabitants of a given region remained Roman Catholic or became Protestant according to the choice of their temporal ruler, the matter of whether a person or a region is Roman Catholic or Protestant has had the greatest effect not only on such subjective factors as culture and personal attitudes but also on the entire social and political profile of Germany. The very name of the numerically largest political party, the Christian Democratic Union, attests to the fact that it is a federation of an older middle-class Roman Catholic party with its Protestant counterparts. The Social Democratic Party, by contrast, owed much of its early orientation not only to Marxism but also to antisectarianism.

The overwhelmingly Protestant areas of the Reich lie in eastern Germany; the balance between Protestant and Roman Catholic in the former western sector is almost equal, with the Roman Catholic population in numerical ascendancy. A minute fraction of the German public belongs to what are known as the free churches, such as Evangelical Methodists, Calvinists, Old Catholics, Jehovah's Witnesses, and Eastern Orthodox. The remainder, those people professing no religion (*Konfessionslose*), have been sharply on the increase, not only because of the secularizing influence of socialism in the East and a slackening of religious loyalties in general but also as a means of avoiding the Federal Republic's "church tax." That relic of Bismarckian Germany requires a citizen to pay a certain percentage of his net income tax to whatever church he acknowledges, despite the fact that, strictly speaking, there is no established state church.

The predominantly Roman Catholic districts are found in the Rhineland and Bavaria. The Rhine-Main, Rhineland-Palatinate, and southwestern areas are mixed, while the northwest is overwhelmingly Protestant. In the five eastern *Länder* and Berlin, Protestants outnumber Roman Catholics 7 to 1. Under communist rule, freedom of religious choice was protected by the constitution, but religious adherence declined, especially among the young, who were encouraged to substitute a secular ceremony for religious confirmation. Protestant (Lutheran) churches, however, acted as rallying points for supporters of unofficial protest groups, leading ultimately to the demonstrations that brought down the communist government in 1989. Prior to that event, the government had restored many historic churches and the Berlin Great Synagogue, while support from West German Christians had allowed the construction of some churches and social facilities in new housing areas. Most of Germany's Jews are formally attached to Jewish communities. Among the *Gastarbeiter,* most of the Muslims are from Turkey.

Lutheran church

DEMOGRAPHIC TRENDS

The Federal Republic has one of the world's lowest birth rates. Life expectancy is increasing, but deaths outnumber births each year.

During the partition, the two Germanys experienced markedly different demographic developments. In the decades following World War II, West Germany absorbed some 15 million expellees and escapees from the eastern territories of the former Reich, ethnic Germans from non-German countries of eastern Europe, and displaced persons of non-German origin. In addition, foreign residents came to constitute approximately 8 percent of the population, accounting for 65 percent of the population growth from the 1950s through the 1980s.

In East Germany, the total population declined from its immediate postwar maximum of 17 million. War losses left a disproportionate number of women (still 112 women to 100 men in the early 1980s) and of the elderly. Population distribution was further altered by the migration of many young, often highly trained East Germans to West Germany prior to the building of the Berlin Wall in 1961. A period of relative demographic tranquility and some structural improvement ended with the discovery in late summer 1989 of escape routes to West Germany by way of liberalized Hungary and Czechoslovakia; the ensuing flood quickly forced the breaching of the Berlin Wall and of the western frontier defenses. An estimated 333,000 people left for the West in 1989, and the flow continued at a rate of 2,000 to 3,000 per day into 1990. These were, once again, from the better-educated and better-skilled elements of the population: doctors, nurses, and industrial workers. Their addition to West Germany's population, and the consequent claims on housing and social services, met with increasing local resentment; their subtraction from East Germany's population was a devastating blow to a moribund economy. Reunification, first of the economy and then of the German nation itself, did not solve these problems but did offer a political and fiscal framework for their solution.

The *Gastarbeiter*

Exclusive of foreign military personnel and their families, approximately three-quarters of all foreigners residing in the former western sector are workers who have migrated from the Mediterranean countries—chiefly Turkey, Yugoslavia, Italy, and Greece—to help relieve the chronic labour shortage. Although the length of stay for most of

these *Gastarbeiter* is intended to be of limited duration, they form a sizable if constantly changing minority. *Gastarbeiter* often bring their families to the Federal Republic or marry during their stay, but with the ultimate goal of returning to their homeland. Yet many families have permanently established themselves, operate their own businesses, and are in the second or even third generation of residence. Under German law, however, neither length of residence nor the fact of birth on German soil confers the right of citizenship to non-Germans.

Another element of the foreign population are the highly educated or skilled migrants from within the European Communities, who are allowed to migrate freely and work in any member country without a work permit. Attracted to the Federal Republic by job opportunities or higher wages, these foreigners retain their original nationality and mostly return home. Another group, of particular significance following the events of 1989, are the so-called *Aussiedler* ("out-settlers"), ethnic Germans or other migrants from eastern Europe with blood or kinship ties with Germany, who are permitted to migrate legally from nations of the former socialist bloc.

Two trends can be noted among the native German population. First, younger people are more mobile than their elders and more willing than previous generations to move about the country as their work requires. Second, the population is shifting to the south, where better work opportunities in high-technology and service industries combine with attractive leisure amenities and proximity to the recreation areas of the Alps and the Mediterranean.

POPULATION DISTRIBUTION

Various historical factors, including contrasting government policies toward agriculture and industry, gave rise to differences in population distribution between the two Germanys. While about half of the former western sector's population lives in or near the major concentrations of population and the distribution of population with respect to habitable land is uneven, western Germany nevertheless is not dominated by any particular region or city. There is no one magnetic metropolis that dominates the landscape. Economic opportunity in the western parts is not confined to the larger centres of population and industry. The spread of light industry to previously sparsely populated and underdeveloped areas, together with the government's policy of continually improving its already excellent transportation facilities, is expected to improve the country's comparatively even distribution of its dense population.

In the east, the highest population densities are in Berlin and in the industrialized southern districts of Chemnitz, Halle, Leipzig, and Dresden. The lowest population densities are in the northern lowlands. Rural areas experienced massive out-migration at the time of the imposition of state and cooperative farming and flight to the West before 1961. Despite above-average rates of natural increase in rural areas, population decline through migration continued. There has also been a process of concentration of rural population into larger settlements. Early gains from migration were experienced in areas of planned industrial development (Eisenhüttenstadt, Rostock, Schwedt, Hoyerswerda). After initial postwar recovery the cities of the south lost population, reflecting industrial decline, rationalization of production, and unattractive environments. East Berlin and its satellite towns were the principal targets of migration until the attractions of the Federal Republic became accessible in 1989.

The economy

Germany is one of the world's greatest economic powers. In the late 1980s West Germany alone ranked fourth in terms of gross national product, of which it exported one-third. The accession of East Germany in 1990, an enormous economic burden in the short run, offered the potential for unprecedented growth. The strength of the economy derives principally from its manufacturing. It is a social market economy, essentially a free market economy with a clear consensus of the roles of government, management, labour, and finance in the marketplace.

Social market economy

ADMINISTRATION OF THE ECONOMY

With a guarantee under the Basic Law (*Grundgesetz*) of free enterprise and private property, the government exercises a mainly regulatory function. Government policy is to intervene as little as possible in matters such as the setting of wages and prices, but to set the terms under which the market operates—for example, in maintaining vigilance against unfair competition and monopolies. The state takes a direct hand in goods and services essential to the economy and to society but in which no profit is to be made, notably transportation, post and telecommunications, hard-coal mining, and agriculture.

The principle that profit must defer to utility also applies in matters of labour and social policy. The government fosters free collective bargaining and the perception that there are two bargaining partners, called "social partners," in labour-management agreements: the respective trade unions and management associations. Labour is organized along industrywide lines; a trade union represents an entire branch of industry, not simply a single trade or skill. Representatives of labour serve on managerial boards, and workers increasingly may acquire stock in their own companies.

The government's hand in the social sector is measured in terms of what is needed to ensure the greatest social benefit while balancing individual responsibility and incentive with the assurance of basic needs and protections. But benefits—including health protection, unemployment and disability compensation, maternity and child-care provisions, job retraining, pensions, and many others—derive from a spirit of shared responsibility and are paid by contributions from individuals, employers, and public funds. The worker pays taxes and contributions in the knowledge that assistance and protection will be there when needed.

It is not the state alone that is called upon to achieve the goals of stable prices, full employment, adequate growth, and balanced foreign trade. The independent Deutsche Bundesbank ("German Federal Bank"), employer associations, and trade unions are also called in on the decision making. The Economic Policy Council, composed of the federal ministers of economics and finance, one delegate apiece from each *Land* government, and representatives from the communes (municipalities), with the participation of the Federal Bank, serves to achieve a unitary approach to fiscal policy, while the Financial Planning Council of similar composition coordinates federal, *Land,* and local financial policies. The federal and *Land* governments are committed to financial planning several years in advance to anticipate needs and changes in the national economy. A Board of Experts for the Assessment of Overall Economic Trends, set up in 1963 and known as the "five wise men," produces an evaluation of overall economic developments each year to assist in national economic decision making. The federal government submits an annual economic report to the Bundestag and Bundesrat that contains a response to the annual evaluation of the Board of Experts and an outline of the economic and financial policies it is pursuing.

Organs of economic planning

The major source of revenue for all levels of government in the Federal Republic is taxes. The federal government takes slightly less than half of total revenues, and the *Land* and local governments share the remainder. Five types of taxes account for four-fifths of all revenues: income, corporation, turnover, crude oil, and trade. Some taxes—wage (the largest single source of revenue), income, capital gains, corporation, and turnover (including import turnover)—are shared by the three levels of government. Others devolve only to the federal government, are passed on as contributions to the European Communities, or are reserved for the *Länder* governments. Local governments can derive taxes directly from land (real estate), trade, and certain types of consumption and expenditure. Since the taxing potential of the *Länder* is unevenly distributed, the economically smaller or weaker *Länder* share in the tax revenue of the richer or more populous *Länder* through a process of "horizontal financial equalization."

As one of the world's richest countries, the Federal Republic has long concerned itself with the complex problem of assistance to developing countries. To promote stability

in Third World nations through helping them develop their basic needs, economies, and infrastructure is a major aim of government policy. The structure of the Federal Republic's foreign aid program is highly complex and is effected at the state level through bilateral and multilateral cooperation, especially through agencies of the United Nations, and by debt remission. At the private level a large number of public service organizations, foundations, trade unions, and the like are directly engaged in cooperation and assistance programs that range from agriculture, health, and education to government administration.

THE FORMER EAST GERMAN SYSTEM

Prior to the collapse of the government in 1989, the economy of the German Democratic Republic was explicitly socialist. Ownership of virtually all the means of production and distribution was placed after 1949 in the hands of the people, as mediated by the Socialist Unity Party. Unlike the typical Soviet-style states of eastern Europe, the German Democratic Republic had previously been part of an advanced economy of the Western capitalist type. Its surviving industrial equipment and infrastructure, inherited skills, and high level of scientific and technical education enabled it to develop the economy and advance the standard of living to a level markedly higher than those of the other socialist countries, though well below those of western Europe.

Central planning. All decisions on the main directions of the economy were taken by the Politburo or the Secretariat of the Central Committee of the Socialist Unity Party. Policies were laid down in advance for an entire planning period, usually of five years, and, even though the plans were broken down into successive one-year slices, the economic system proved to be an inflexible one, reacting reluctantly to experience.

Economic planning organizations

The German Democratic Republic, with its reliance upon exports to pay for imported raw materials, had a growing awareness that its manufactures were being outclassed in world markets by the rapid rate of technological innovation characteristic of capitalist economies. The government's reaction was to attempt to switch to more sophisticated production methods and to more advanced products in fields such as electronics. A key mechanism in the intended industrial transformation was the introduction in 1979–80 of the policy of concentrating groups of enterprises into large nationally owned combines, formed around a core provided by the leading firm in the branch of industry concerned, with perhaps 25,000 workers under a single managing director.

Wage differentials under the communist regime tended to be low within the predominant nationalized sector; power and influence were often more significant than money. Trade unions were not expected to fight for better pay but to collaborate in achieving the party's economic objectives; they had important functions in relation to safety, welfare, and leisure programs and in administering the social security system for their members. During the transition period prior to reunification, trade unions based on the Western model took over in East Germany, and social insurance was aligned with the West German system.

Throughout its history the German Democratic Republic suffered from an acute shortage of labour, reflecting both the unbalanced age structure of the population and the losses through flight to the Federal Republic. As in other European countries, an additional problem from the late 1980s was the reduced number of young persons entering the labour pool. The labour shortage in the German Democratic Republic was also increased by the need to staff the complex and overlapping party, state, and trade union bureaucracies, the armed forces, and the various police forces. Measures to meet the shortage included a system of rationing labour in relation to perceived national needs and the assumption that all women should be in full-time employment. After the opening of the Berlin Wall and of the western boundary in late 1989, with the consequent population outflow, the shortage of workers in industry and the public services became even more acute.

Industrial deficiencies

The collapse of the socialist administration revealed gross industrial deficiencies that had been concealed behind a screen of secrecy and falsified statistics. With transformation to a market economy, the large combines were taken out of ministerial control and transformed into limited companies (stock corporations). They were burdened with debt, however, because the former government had appropriated their profits and required them to borrow their capital. Few East German enterprises were the equal of Western firms in technology or productivity, while their workers, lacking financial incentives and frequently obliged to await the belated delivery of components, had become accustomed to a leisurely pace of activity. Many industrial facilities were fit only for the scrap heap, with a prospect of unemployment that was, at least officially, unprecedented. Yet the work force was basically well-trained and low-paid, making the eastern sector attractive for takeovers and joint production projects by West German firms; the automobile industry (for example, Volkswagenwerk AG) was one of the first to move in.

A particular problem was that the former government, in its drive for production at any price, had neglected environmental considerations. West Germany and West Berlin had long been recipients of air and water pollution from the East. Many chemical plants in the southeast (with attendant settlements) were so dangerous as to require total abandonment. Reunified Germany was likely to try to restrict the burning of untreated brown coal (lignite) as a source of energy, because of the extent of atmospheric pollution involved. The discharge of waste salt from the potash industry into rivers was also of great concern. Safety standards at East German nuclear power stations, built on the Soviet model, were suspect, while the areas in the south of the country where uranium had been mined for the benefit of the Soviet Union were believed to be contaminated by radiation.

Distribution and services. Retail distribution during the socialist era was almost totally in the hands of state-owned chains. Supplies of food and basic consumer goods were generally ample, although lacking in variety and beset by periodic shortages; their prices were heavily subsidized. By contrast, goods deemed to be luxuries were high-priced and difficult to obtain; delivery time for a new automobile, for example, could be eight or more years.

The centralized provision of services through nationalized concerns and local administrations was a generally recognized weakness. Accordingly, persons with a skill—such as medical practitioners, building craftsmen, or auto mechanics—were able to offer their services on a barter basis or in return for illicitly held foreign currency. Retail facilities, the range of goods available for sale, medical provisions, and cultural facilities were unevenly distributed; the capital, as the showcase for visitors, was the best-provided, with the small towns and the countryside coming last. All these conditions of retail distribution and services were modified by the transfer to the deutsche mark currency in 1990 and by increasing privatization.

Domestic construction in the 1970s and '80s featured the erection of vast complexes of high-rise apartment blocks from standardized prefabricated elements. The result was monotonous, but people were housed under modern conditions. There was some building of small private homes on the urban fringe or even of holiday cottages (dachas), but building materials, whether for new construction or for repair, were difficult for private individuals to obtain.

Foreign tourism in the German Democratic Republic was centrally controlled by a government office and mainly took the form of organized group tours of East Berlin, Dresden, and other cities. Citizens had only limited possibilities for visits to other socialist countries before the changes of 1989.

Tourism

Finance. During the 40 years of communist rule, economic nationalization swept away most of the financial institutions of the former capitalist system. Instead there was a unified state finance system, the object of which was to facilitate the implementation of state policies. Under this system, the state budget of the German Democratic Republic covered a much wider field than would that of a capitalist country. In addition to the expenditures of the central administration, it included those of local government at its various levels. It embraced the banks and

insurance institutions, directing them in the employment of their funds. Above all, the state had to meet most of the capital needs and running costs of the economy, for both state-owned concerns and cooperatives. Tax on personal income was not an important budget element, but there were heavy sales taxes on luxury goods such as television sets. The bulk of the state's income came from publicly owned enterprises in the form of profits, charges on capital, and a payroll tax.

All this changed with the first formal act of reunion, the state treaty introducing the West German deutsche mark as a common currency on July 1, 1990. The treaty had associated provisions to protect East German land and property from speculators and to ease the transition of East German industry and agriculture to the market economy. Already Western banks had begun to move back into the Eastern sector. In May 1990 the Federal Republic agreed on massive loans to help bail out the East German economy.

Trade. In the former East Germany machinery of all kinds, increasingly combined with electronic controls, made up by far the largest category of exports. Imports were mainly of raw materials—such as hard coal, oil, and natural gas—or of semifinished products, such as rolled steel. Machines and vehicles not made domestically were also imported.

Under the socialist government about two-thirds of foreign trade was conducted with members of the Eastern-bloc Council for Mutual Economic Assistance (Comecon), principally the Soviet Union, and more than a quarter with Western countries, led by West Germany, to which market the Democratic Republic had unrestricted access. The country's extensive web of trade agreements with the Eastern bloc posed a considerable problem to the negotiators of reunification. On the other hand, East Germany, unlike the other former Soviet-bloc states, automatically became part of the European Community by uniting with the Federal Republic.

Ownership. One of the thorniest economic questions attending reunification was that of ownership of property. The former communist government took into public ownership most of industry and distribution, much housing, and nearly all agricultural land. Former owners or their heirs might expect to resume their rights under the laws of the Federal Republic. Yet more than 40 years had passed since nationalization, and the position was complex. Much state investment had gone into the industrial plants. It was scarcely possible to envisage the reconstitution of the original pattern of small agricultural holdings. Whole urban quarters, industrial plants, and public installations of all kinds had been built on land that might be held to belong to private citizens or even to foreigners. Under terms of the unification treaty signed by representatives from the two Germanys on Aug. 31, 1990, current ownership was protected and compensation in any land disputes was guaranteed by the government.

RESOURCES

Germany has a relatively low endowment in natural resources, and as a major industrial country it must import most of its raw materials. It has enough anthracite and lignite coal for its own needs. Hard coal deposits are found mainly in the Rhine-Ruhr region and the Saarland; brown coal is found near the Harz Mountains, near Cologne, along the Polish frontier, and around Halle and Leipzig. Immediately prior to reunification, East Germany was the world's largest producer of lignite, and heavy use of the fuel was a major factor in that region's critical pollution problem. The use of coal has declined as a major energy source in the former western sector, deferring to petroleum, natural gas, and nuclear power. Other minerals found in economic abundance are salt and potash, mined north and south of the Harz Mountains. The uranium ores of the Erzgebirge, in the former eastern sector, were worked after World War II for the benefit of the Soviet Union. Materials needed for the construction industry are plentiful. The water supply poses a difficult problem in the urbanized, industrialized, and polluted east. Major dams and reservoirs are concentrated in the uplands of the south and west, while in the sandy glacial outwash areas to the north much water is drawn from wells and from lakes or rivers.

Coal deposits

Some petroleum deposits and natural gas deposits are located on German soil, mainly along the North Sea littoral, especially in the Emsland in the northwest border region with The Netherlands. While domestic crude oil production can supply only a tiny percentage of the country's needs, its natural gas reserves supply one-third (the rest is imported from The Netherlands, Norway, and the Soviet Union). Thermal (coal-powered) electric plants, nuclear power stations, and hydroelectric plants furnish the nation's electricity.

AGRICULTURE, FORESTRY, AND FISHING

Agriculture. As in the other economic sectors, differing national policies during the partition created divergences in agricultural development between the two Germanys. The former western sector is still basically a region of peasant family farms. Only 5 percent of the entire work force is engaged in farming, but 95 percent of all farms are smaller than 125 acres. About half of the smaller farms are part-time operations, the family or owner also deriving income from another source.

In the east, by contrast, following conquest by the Soviet army at the end of World War II, many of the large estates that had been characteristic of the region were split up or retained as state farms. From 1952 to 1960 virtually all small farms in the Democratic Republic were united, under strong political pressure, to form agricultural cooperatives. Agricultural production was increasingly concentrated into specialized units of extremely large size; by the mid-1980s state or cooperative crop-producing enterprises averaged 11,292 acres (4,570 hectares). The increased scale of production, modern machinery, and technological innovation led to increased production, despite a marked decrease of workers. Among the formidable challenges of the reunification process was the reconciliation of the large-scale industrial agriculture of East Germany with the generally small-farm structure of the western sector.

The agricultural cooperatives

In areas of high natural fertility wheat, barley, corn (maize), and sugar beets are Germany's principal crops. The poorer soils of the outwash plains in the north and of the uplands grow corn, rye, oats, green fodder, and potatoes. Livestock includes pigs, cattle, sheep, horses, ducks, and other fowl. The primary dairylands are in the northern coastal lowlands and in Bavaria. Crops and their traditional areas of specialization include the fruits of the Havel lake country near Potsdam, the Halle region, and Thuringia and green vegetables of the peaty soils of the Spreewald. Vineyards are mainly in the west, especially in or near the valleys of the Rhine, Moselle, Main, and Neckar rivers, although the hill slopes of the Elbe valley near Dresden also produce wine grapes.

At the time of reunification, West Germany produced 80 percent of its food requirements, and increased productivity and guaranteed prices resulted in vast surpluses (especially of butter, meat, wheat, and wine), the economic disposal of which constituted a classic embarrassment of riches.

Forestry. Approximately one-third of German territory is covered by woodlands, mainly conifers. The most densely forested districts are in the *Länder* of Hesse, Rhineland-Palatinate, Baden-Württemberg, Bavaria, Thuringia, and Saxony. The federal government, *Länder,* and municipalities own more than half of the forested land; the remainder is private property. Forests in the Federal Republic are valued not only as a source of timber and an environmental asset but also as recreational areas. Atmospheric pollution, especially from dense motor-vehicle traffic and industrial and power-station exhaust, now poses an immediate and tangible threat to the country's forests. In the Democratic Republic, forests were mostly held in public ownership. Conifers predominate, notably pine (*Pinus sylvestris*), spruce, and fir. The forests have largely recovered from excessive cutting during and after World War II, and they now make a substantial contribution to requirements for lumber, pulp, and board. Pine resin is delivered to the chemical industry.



Harvesting potatoes, Brandenburg, Ger.
ADN Zentralbild

Fishing. The German fishing fleet met with drastic reductions in the 1970s and '80s. The decimation of traditional fishing waters, together with a reduction in the size of fishing areas because of the extension of national zones by coastal countries to 200 nautical miles, caused grave declines in the catch of cod, sea salmon, redfish, and herring from the high seas and coastal waters. Baltic Sea fishing (traditionally for herring) was reorganized prior to reunification, based on a processing plant at Sassnitz, on the island of Rügen. Of greater significance, however, is the deep-sea fishing fleet, based on another processing plant in the extended Rostock harbour.

INDUSTRY

The principal industries that make up the economic might of Germany are motor-vehicle production, mechanical engineering, electrical engineering and electronics, chemicals, and food processing. About half of all industries remain small, employing fewer than 50 people. Only a very small percentage of western Germany's industrial companies can be described as large, with 500 or more employees; they are in the mining, steel, petroleum-processing, and motor-vehicle sectors.

The Federal Republic is the third largest producer of automobiles in the world after the United States and Japan and exports 60 percent of the automobiles it produces. Mechanical engineering firms, which account for a large portion of the smaller enterprises, and the electrical engineering and electronics industry are also export-oriented. The country's high-technology firms, which also play a major role domestically in innovation, modernization, and improving efficiency, tend to be located in Baden-Württemberg and Bavaria. As a major source of new jobs, they are a factor in the population shift to the south.

Traditional industries

The chemical industry is the most important of the traditional basic-materials industries. Less successful have been the mining and iron and steel industries. While the shipbuilding industry has had to face tough competition from Japan and South Korea, it is of vital importance in the economy of the maritime north.

Other important though smaller industries are aerospace, precision engineering, optics, and watch- and clockmaking. Of considerable importance in the nonmanufacturing sector is tourism, with Germany ranking among the world's most popular nations as a tourist destination.

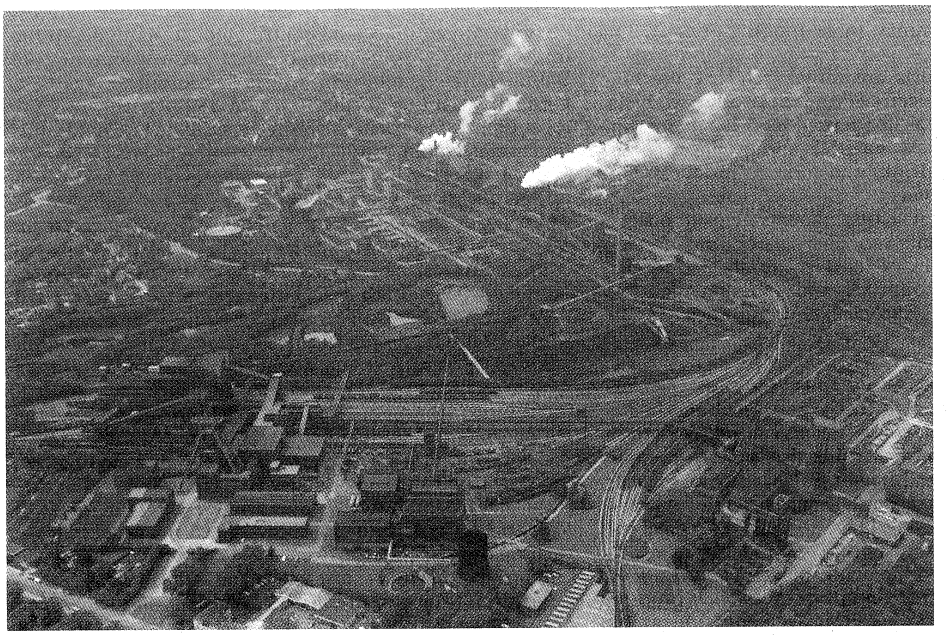
Under the socialist regime in East Germany, the environmental and health consequences of industrial pollution were virtually ignored for 40 years. Over the same period, facilities and equipment suffered an alarming degree of deterioration. As a result, the entire industrial complex of the Democratic Republic—the showpiece of the Eastern bloc—was, at the time of reunification, expected to undergo a massive assessment and reorganization, with shutdown of some major enterprises considered inevitable.

As discussed above, a major contributing factor to the pollution problem was the eastern sector's traditional reliance on lignite as its main energy resource. The mineral, also known as brown coal or wood coal, was mined from thick seams in vast open pits, primarily located in the Halle-Leipzig area and in the Lausitz field east of the Elbe. The coal was commonly compressed into briquettes for domestic use, but about half was fed directly into thermal power stations. The only other significant source of electricity was atomic power; the plants were located in the north, to serve areas remote from the lignite fields. Coal gas was produced at large plants in the lignite fields and distributed by pipeline. During the partition, the main supplies of natural gas came from the Soviet Union by way of Czechoslovakia, while the principal crude oil supply was by pipeline from the Soviet Union through Poland to Schwedt, which was also reached by pipeline from the port of Rostock. Schwedt was the principal refining centre for East Germany and Berlin, but some of the crude oil was pumped for processing at plants in the Halle-Leipzig area.

Prior to reunification, the chemical industry had shifted its basis from lignite, as both a raw material and a source of energy, to oil and natural gas. The complex of chemical plants in the Halle-Leipzig-Bitterfeld area was supplemented by outlying producers at Schwedt and Lower Lusatia. Further processing of chemical-based products took place at widespread locations. The Democratic Republic was also the world's third largest producer of potash salts.

Metallurgical products in the east included quality steels, aluminum, and other metals. Vehicle production included electric locomotives, rail freight cars, small automobiles, and light trucks. Machinery production vied with chemicals for the leading position in manufacturing. Major centres were the southern industrial towns and the East Berlin region. Shipbuilding greatly expanded in the ports of the Baltic coast.

The German Democratic Republic traditionally specialized in the manufacture of electrical equipment; electrical equipment and electronics branches had the highest growth rate of all industrial branches from 1960 to 1989. East Berlin was the leading centre of production, followed by Dresden and other cities in the south. The traditional optical and precision-instrument industries (Zeiss Jena) also greatly expanded and varied their range.



City of Bochum in the Ruhr industrial district, Nordrhein-Westfalen, Ger.
© Georg Gerster—Comstock, Inc.

FINANCE

The banking system in the Federal Republic is composed of the central banking system, the Deutsche Bundesbank and its main offices in each of the *Länder*—the *Land* Central Banks—and the private banking sector, consisting of private commercial banks, savings and credit institutions, and cooperative credit institutions.

The central banking system. The Deutsche Bundesbank, with headquarters in Frankfurt am Main, the city considered to be the financial centre of the Federal Republic, issues bank notes, regulates the circulation of bank notes and coins, and controls the credit supply. While it is independent of the federal government, the Deutsche Bundesbank is expected to support the government's general economic policy. It controls monetary supply policy; members of the federal government may participate in its deliberations, but they have no vote.

On July 1, 1990, the Deutsche Bundesbank effected the conversion of the East German monetary system to that of the Federal Republic. Upon political reunification, the all-German government assumed the debt of the former eastern sector, with the provision that the five new eastern *Länder* would repay half of the amount of that debt. While the former West German *Länder* shared proceeds from sales taxes on a need basis, the new states were to be considered separately so as to prevent a disproportionate drain on revenues.

The private banking sector. The private commercial banks, of which there are several hundred, are headed in size by the "Big Three": the Deutsche Bank Aktiengesellschaft, or AG ("Limited-Liability Company"), the Dresdner Bank AG, and the Commerzbank AG, which have thousands of branches throughout the Federal Republic. The commercial banks conduct the customary full range of banking business; they provide long-term finance to industry, the government, and local authorities, and they sell stocks. German private banks may also participate directly in the industrial sector; the largest bank, Deutsche Bank Aktiengesellschaft, is the major shareholder in Germany's largest company, Daimler-Benz AG.

Credit and investment. Credit institutions under public law are the savings banks and the *Land* banks, which act as central banks and clearinghouses for the savings banks. Their scope of activity is limited to credits, investments, and money transfers, with the focus on home financing. The central giro institutions act as clearinghouses for the savings banks. The state-owned Kreditanstalt für Wiederaufbau ("Reconstruction Loan Corporation") is the channel for public aid to developing countries.

The cooperative banks are headed by the DG Bank (Deutsche Genossenschaftsbank; "German Cooperative Bank"), overseeing several thousand industrial and agricultural credit cooperatives. Germany's fourth largest bank is the Bank für Gemeinwirtschaft ("Bank for Collective Farming"), the cooperative of the German Federation of Labour, the major trade-union organization. There are also public and private mortgage banks, installment credit institutions, and the postal check and postal savings systems operated by the federal postal service.

Stock exchanges

Stock exchanges are located in Bremen, Düsseldorf (the second largest centre of banking), Frankfurt am Main, Hamburg, Hannover, Stuttgart, Munich, and Berlin. Public ownership of stock has tended to remain in the hands of seasoned investors.

Of German investments overseas, 80 percent is invested in Western industrial countries, chiefly in neighbouring countries of Europe and in the United States. Most of the remainder is invested in developing countries as a major adjunct of the Federal Republic's foreign aid policies. Such investment serves to provide the long-term capital and technical and managerial expertise required in the developing countries.

TRADE

In 1986 the former West Germany replaced the United States as the world's leading trading nation. While this was largely due to a sudden drop in the value of the dollar, it put the great strength of the German economy into clear perspective. After 1952 the country showed a surplus of exports over imports and an increase in the value of those exports by the mid-1980s of more than 150 percent.

Germany's most important trading partners are its fellow members of the European Communities, followed by the United States. The Federal Republic receives most of its imports from The Netherlands, France, Italy, the United Kingdom, Belgium and Luxembourg, and the United States. The major recipients of German exports are France, the United States, The Netherlands, the United Kingdom, Italy, and Belgium and Luxembourg. Switzerland, Austria, Sweden, the Comecon nations, Japan, and the Soviet Union are also important trading partners.

Intra-German commerce

During the era of partition, the Bonn government considered the Federal Republic and the German Democratic Republic to be two independent states within one German nation. Accordingly, all trade with East Germany was regarded as intra-German commerce, hence not "foreign" trade. Following the collapse of the communist regime in the East, economic union became actual in July 1990,

when East Germany adopted West Germany's deutsche mark as its unit of currency three months prior to full political reunification.

Germany's major exports are motor vehicles, machinery of all descriptions, electrical and electronic products, and chemicals, as well as food products. Its major imports are capital goods; raw materials and producer goods; consumer goods; food, drink, and tobacco products; and mining and energy products.

Germany's positive balance of payments helps to offset a deficit in foreign payments caused by its citizens' foreign travel, by the remittances of its large number of foreign workers to their home countries, and by its payments to the various bodies of the European Communities.

TRANSPORTATION

The Federal Republic not only must serve its own extremely complex needs as a densely populated and highly industrialized nation, but, because of its central location in Europe, it must also function as Europe's principal nation of transit for passenger and freight traffic. In consequence, its system of overland transport—highways, railways, and waterways—and, to a lesser but adequate degree, of air transport is among the most intricate and highly developed of any nation. Elaborate plans, made in coordination with neighbouring countries and partners in the European Communities, call for further improvements and ambitious long-range programs to expand and improve existing facilities and to incorporate the most advanced concepts in transport engineering to handle the growing volume of traffic anticipated in the decades ahead. More limited-access motorways and widened highways will lessen the present immense burden on Germany's labyrinthine road system, while the continuing modernization of existing rail facilities will be augmented by the construction of new lines to relieve the most heavily traveled routes.

Patterns of traffic movement. Although throughout history Germany has served as a crossroads for European traffic, both from north to south and from east to west, the geographic contours of Germany and its inherited overland routes, as well as formerly restricted political borders, have made the north-south and northwest-southeast routes of greater significance than those running from east to west. The emergence of democracy in eastern Europe—and particularly German reunification—made possible the revival of east-west transportation routes.

Predominance of north-south routes

The notable exceptions to these patterns during the era of partition were the freight and passenger routes to West Berlin; access to the city, except for two of the three air corridors and the Elbe River, was restricted to certain rail and autobahn routes. West Germany had to pay transit dues in hard currency and contributed substantially to the improvement of routes, notably to the construction of the Berlin-Hamburg autobahn, which was completed in 1982. Until the partition of the Reich in 1945, Berlin was the focal point of the nation's railway system, but today this function is all but vestigial, even within eastern Germany.

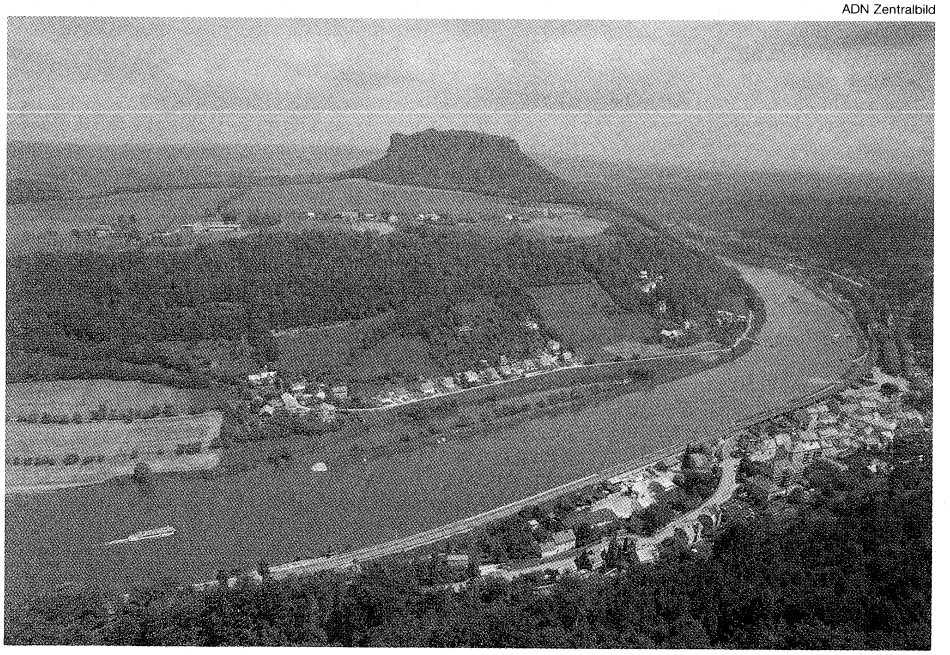
National routes. Germany's major overland routes follow parallel patterns in both the highway and rail systems. From north to south, traffic gathering from the direction of Scandinavia focuses at Hamburg. It then follows a course southeastward to Berlin and southward to Munich, or roughly due south via Hannover and Würzburg and then eastward toward Nürnberg or southeastward toward Munich. Alternatively, it may go from Hamburg southwestward toward Bremen, whence it veers south and then west to the Rhine-Ruhr district, from which area the major routes of transport follow the Rhine valley, past the Rhine-Main district to all points south.

Major intermediary routes are the rail and highway links from Rostock, Hamburg, or Hannover to Berlin; from Berlin southward to Leipzig or Dresden and on to Nürnberg and Munich; from Hamburg and Hannover to Frankfurt am Main; from the Rhine-Ruhr district to Hannover; from Frankfurt am Main to Würzburg, Nürnberg, and Munich; from the Mannheim-Ludwigshafen area westward to the Saar and northern France; and from the Mannheim-Heidelberg area to Karlsruhe or, farther eastward, to Stuttgart, Ulm, Augsburg, and Munich. Munich is the great highway and rail terminal point for traffic to and from Austria, Italy, and the Balkans. The major routes to southwestern Europe run south past Karlsruhe and up the Rhine valley—with a major access to eastern France at Strasbourg—or south past Freiburg im Breisgau to the juncture of the German, Swiss, and French borders at Basel.

Germany's unexcelled system of overland transport is complemented by its great endowment of navigable rivers and their interconnections by means of an intricate system of canals. Apart from the Danube River and its tributaries, which flow eastward toward the Balkans but eventually are scheduled to be linked by canal to the northern water routes, these rivers have access to the North Sea and, via a link with the Elbe, to the Baltic Sea. Although three of Germany's major rivers—the Elbe, the Weser, and the Ems—empty into the North Sea, a major portion of freight to or from overseas passes through The Netherlands and Europe's largest port, Rotterdam, for the Rhine route via Rotterdam is closer by almost half the distance to the North Sea and the Atlantic from the Rhine-Ruhr district

Rivers and canals

ADN Zentralbild



The Elbe River, with the town of Königstein (right) and Lilienstein hill (top), Saxony, Ger.

than even the Ems. As a corollary, The Netherlands' extensive motor-freight fleet has been developed to service goods passing through the port of Rotterdam, and a major portion of this fleet is deployed to and in transit through the Federal Republic via North Rhine-Westphalia.

Local and interurban services. Germany's systems of passenger transport are among the most thoroughly developed and efficient in the world. A dense network of mainline railways not only provides frequent and speedy service in long-distance passenger traffic but also serves simultaneously for local commuter traffic.

Local urban transport is no less intensely structured: local bus systems operate in even the smallest towns, and all communities have access to the regional interurban bus lines. German cities, far from abandoning street railway lines in the post-1945 era in favour of municipal bus lines, continue to rely heavily upon urban rail service; and in the larger cities, municipal buses are usually ancillary to the surface tram lines. Although Berlin and Hamburg have maintained both underground and elevated train services for decades and Wuppertal constructed an 8.3-mile-long elevated monorail in 1900, other major German cities did not begin to construct major rapid-transit facilities until relatively late. Most of these rapid-transit systems, in such cities as Frankfurt am Main, Cologne, Stuttgart, and Hannover, represent essentially a submerging of the surface rail lines in the central portions of the city. The exception is Munich, which built a full-scale underground system independent of the surface lines, complemented by an independent fast-rail service linking communities within a 25-mile radius of the city centre.

Interurban bus transport, frequently operated by the federal postal service as a relic of the ancient postal coaches, is limited in most instances to localized suburban and nearby interurban services. Because of the superior train service available, long-distance bus runs play a relatively minor role in passenger transport.

Of the major forms of long-distance freight carriage of all kinds, including transshipments, the railways haul a substantial part of all goods, followed by inland shipping, overseas shipping, and motor-freight carriers. The remainder of freight transport is accounted for by pipeline (for chiefly oil and natural gas), plus a relatively minute portion of goods shipped by airfreight.

Highways. Several tens of thousands of miles of roads and highways of all categories remained in Germany after World War II. Even the best of these were antiquated in terms of early postwar traffic demands, and few exceeded two lanes in width. Superimposed on the older highway system was the badly damaged system of limited-access motorways known as the autobahn, begun under the Nazi regime in the mid-1930s and considered the marvel of the world in its time. It was in operation well before comparable superhighways in the United States. But, though the autobahn complex linked the major industrial and population centres, many important segments had not been completed by the war's end, and its relatively narrow dual lanes, impeded by war-damaged or incomplete sections, were soon overtaxed by the mounting volume of traffic. Improvements and extensions of the autobahn network were begun in the early 1950s, and by 1990 it covered more than 5,000 miles. The ring autobahn surrounding Berlin was not completed until 1979; postwar additions in the former eastern sector included reinforcement of the transverse network in the economically important south and new routes northward to Rostock and Hamburg. With similar developments in neighbouring countries, it is now possible to cross over directly without interruption from the autobahn to the motorways of Denmark, Belgium and Luxembourg, France, Switzerland, and Austria.

The autobahn

Germany's love affair with the automobile has placed a severe strain on the country's otherwise excellent roads. The volume of motor traffic in the western sector alone, for example, increased 16 times between 1950 and 1990, with the number of domestic vehicles enhanced by vehicles from the rest of Europe, all competing for road space that did not even double over the same period. Not surprisingly, there is an alarmingly high rate of road accidents, despite strenuous efforts in road safety. Drivers, through their elected representatives, repeatedly have rejected attempts to impose speed limits on autobahn traffic.

Railways. Germany's first train service operated over a five-mile line between Nürnberg and Fürth in 1835. The privately owned railway corporations formed throughout the late 19th century were transferred to public ownership through the individual states, and by 1920 the entire network was owned and operated by one national corporation, the Deutsche Reichsbahn ("German National Railway"). Some minor lines—for the most part narrow-gauge, resort, or excursion lines—survived under private ownership. Prior to territorial annexations from Austria and Czechoslovakia before the outbreak of World War II, Germany had some 35,000 miles of railway lines. About 18,000 miles of this network, much of it heavily damaged during the war, devolved to West Germany, which by 1951 had taken over operation of service from the occupying powers under the name of the Deutsche Bundesbahn ("German Federal Railway"). Because widespread wartime destruction called for a complete overhaul of capital stock and equipment, the railway system had abandoned all steam locomotives by 1977 in favour of diesel or electric engines.

The East German rail network was extensively rebuilt and modernized in the postwar decades but was reoriented to emphasize links with Comecon neighbours and with the port of Rostock. Many links across the boundary with West Germany were abolished or reduced to single tracks. Internal lines converged on a ring line enabling the avoidance of West Berlin. Within Berlin, the trains, buses, and trams of the public transport were totally divided. An undoubtedly lengthy process of restoring a unified system, both within Berlin and nationally, began in late 1989.

Frequent service is offered by conventional express trains (D-Züge), the somewhat faster Inter City (IC) trains that run between major cities, and the high-speed international TEE (Trans-European Express) trains. On routes covering 275 miles of track, speeds of 125 miles per hour are possible. Two entirely new routes under construction between Hannover and Würzburg and between Mannheim and Stuttgart will permit speeds up to 155 miles per hour. Ultimately, a line is planned from Mannheim to Paris and from Paris via Brussels to Cologne to connect with the French superspeed TGV (Trains de Grande Vitesse) system. Work is also in progress in Berlin on superspeed monorail lines and on a train powered by linear acceleration.

Express trains

Waterways. Germany's intricate system of inland waterways, of which the Rhine is the very symbol, assumes a vital economic importance in the inexpensive haulage of bulk goods, raw materials, and fuels. Of the 4,200 miles of navigable waterways, about two-thirds are represented by rivers and most of the rest by canals, together with a few lake routes, especially on Lake Constance. The Rhine is the major internal water transport route, where the density of ships, barges, freighters, and other vessels from the fleets of five nations is comparable to that of a heavily traveled highway. The Elbe, Havel, and Oder rivers also provide south-north routes, which are linked by east-west canals. Because the maritime outlets of this system—at Hamburg and at Szczecin (Stettin), in Poland—were outside the former Democratic Republic, the government developed the Warnow estuary, downstream from Rostock on the Baltic Sea, into a deep-sea port; imports of bulk raw materials greatly predominated over exports. Some use was still made of Hamburg, which now may resume its role as the major outlet for the Elbe basin. Rostock-Warnemünde is linked by rail ferry to Gedser, Den.; Sassnitz, on the island of Rügen, with Trelleborg, Swed.; and Mukran, on Rügen, with Klaipėda (formerly Memel), Lithuanian S.S.R.

The 200-mile-long Mittelland Canal, which connects Berlin with the Elbe, extends westward across the former inter-German frontier into Lower Saxony and the Greater Hannover industrial region, leads to the Weser River at Minden, and continues westward with the Dortmund-Ems Canal leading to the North Sea port of Emden. The Kiel Canal (also known as the Nord-Ostsee Canal), 61 miles long, has linked the Baltic and the North seas since 1895. Both rivers and canals are constantly being deepened,

channels widened, and locks enlarged to improve this cheap and vital mode of freight haulage. The channelization of the Saar will connect the Saarland to the European waterway system. The most important development in inland shipping for both Germany and all Europe is the linking of the northern river and canal system with that of the Danube and its tributaries and canals via the Main-Danube (Europa) Canal, a project contemplated since the reign of Charlemagne in the 9th century AD. This 106-mile (171-kilometre) trans-European link between the North Sea and the Black Sea was scheduled for completion in 1992. Plans call for a total of 16 locks, with a standard length of 623 feet and a width of 39 feet. The northern link between the Main River (near Bamberg) and Nürnberg was completed in 1972.

Although having only limited access to the open sea, Germany long has ranked high as a maritime power: its merchant marine was the fifth largest in the world at the outbreak of World War II. As with other leading shipping nations, however, the Federal Republic's position among the merchant fleets of the world is obscured by the intricacies of patterns of ownership and through ship registration under flags of convenience. The major German seaports of Bremen, Bremerhaven, Hamburg, Lübeck, and Rostock compete with the more advantageously situated North Sea ports of Rotterdam and Antwerp by excelling in container operations and the swift transfer of cargo. Hamburg, located 68 miles inland on the Elbe, accounts for the greatest number of arrivals and departures.

Airways. Germany, apart from its many technical contributions to aeronautics, was one of the great pioneers of commercial aviation in the years between World Wars I and II. The original Deutsche Lufthansa AG, founded in 1926, whence the state-owned airline Lufthansa takes its name, was an early leader in operating regular passenger service throughout Europe and by 1936 was flying the North Atlantic route. In 1938 this airline offered nonstop passenger service between Berlin and New York City. Since the reestablishment of German commercial aviation in the mid-1950s, Lufthansa has become one of the leading international carriers.

Role in international aviation

It is, in fact, in international rather than in domestic aviation that Germany's role is important, not the least reason for which is the nation's central location in Europe. Because of the relatively short distances between the cities of Germany and the availability of fast and less expensive train service, domestic air service carries a comparatively small share of German travelers. Before the opening of the East German border, however, air service provided a means of travel to and from West Berlin that was free of border formalities. Air freight represents only a relatively small portion of the total of long-distance goods hauled in the Federal Republic.

Air transportation in the Democratic Republic, centred on the airport of Berlin-Schönefelde southeast of Berlin, expanded in stages after 1959. The airport is situated on the Berlin ring railroad and can also be reached by the S-Bahn from Berlin. Accordingly, the rail system, rather than internal airlines, was used for feeder purposes. Berlin-Schönefelde served as the main base of Interflug, the Democratic Republic's international airline.

Commercial airports serve the major cities of Germany, of which Frankfurt Airport, popularly called Rhein-Main, is one of Europe's largest in numbers of passengers and total air freight handled. Because more than one million visitors fly in to Frankfurt am Main each year to attend its numerous international trade fairs, an underground railway connection links Rhein-Main airport with the fairgrounds. Major airports are located in Berlin, Bonn-Cologne, Munich, Düsseldorf, Hamburg, and Leipzig.

Administration and social conditions

The structure and authority of Germany's government is derived from the Grundgesetz, or Basic Law, which was signed and proclaimed on May 23, 1949, after formal consent to the formation of the Federal Republic (then comprising 12 *Länder* [10, following the merger of 3 in 1952] and known as West Germany) had been given by the military governments of the Western occupying powers and on the assent of the parliaments of the *Länder* to form the *Bund* (federation). The extension of full sovereignty was achieved only gradually: many powers and prerogatives, including those of direct intervention, were retained by the Western powers and devolved to the Federal Republic only as it was able to grow in economic and political stability and to be integrated into the Western community of nations. The tripartite offices of military governor were replaced upon the creation of the Federal Republic by those of high commissioners, and upon its achievement of full sovereignty on May 5, 1955, the high commissioners became ambassadors accredited to the president of the republic. Article 23 of the Basic Law provided for the eventual accession of the eastern sector and restoration of a united Germany. With the fulfillment of this provision on Oct. 3, 1990 (see below *History: Germany reunited*), the Basic Law was amended accordingly, thus precluding any claims to former German territories now lying within Poland or the Soviet Union.

In the days of imperial Germany, the society of the nation was among the most intricately hierarchical in all Europe. The social upheaval of two major wars and economic change, though loosening this rigidity, left the basic class structure intact. German society is far from plagued by class consciousness, but a sense of one's station in life is implicitly understood. Education still commands a greater awe in Germany than in many countries; the professor is held in an esteem incomprehensible to foreigners, and the title "doctor" is an all-but-essential credential for advancement not only in the upper echelons of the professions and the civil service, where a certain erudition is not inappropriate, but also—even more baffling to outsiders—in the ranks of business. The greatest loss of prestige has been suffered by the military; the German officer, who in the imperial era could snub bank directors with impunity, now commands mere apathy from the public. Of especially compromised status are the officers of the former East German army, who upon national reunification were subordinated to their western counterparts, against whom they had, in effect, borne arms for 40 years. The older authoritarianism still lurks, but it has been enfeebled beyond repair. As in other advanced nations, the basis of true power has passed to the younger technical and managerial meritocracy.

GOVERNMENT

The constitutional framework

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

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

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

Ministries of the Cabinet

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

The chancellor is immediately assisted by his secretaries of state, who administer various aspects of foreign and internal affairs or the conduct of press and information services; they may exercise wide powers of discretion in carrying out the instructions of the chancellor. The Cabinet ministries—apart from the major areas of foreign policy, finance, defense, internal affairs, justice, and commerce—are responsible for such technical and social functions as post and telecommunications (including the post office, the telephone system, and certain aspects of broadcasting); youth, family, women, and health; economic cooperation; education and science; labour and social affairs; and regional planning, construction, and urban development. A ministry peculiar to the German situation was that devoted to intra-German relations, which concerned itself with matters pertaining to the Federal Republic's dealings with the German Democratic Republic. For an account of German reunification, see below *History.*

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

Methods of constituting the parliamentary chambers

The Bundestag, the cornerstone of the German system of government, was enlarged in 1990 following reunification and all-German elections to some 600 members (subject to slight variation), 144 of them from former East German territory. Its delegates are chosen either in general elections held every four years or in special by-elections. In addition to the members elected by each district, a set of *Länder* delegates at large is elected simultaneously both as a means of ensuring stability and continuity of representation by the major parties in the lower chamber and as a corollary to look to the interests of nation, *Land,* party, or bloc of voters at large. In this latter function the delegates at large serve as a counterbalance to the parochial tendencies inherent in strict representation by district constituency mandate.

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

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

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

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

THE FORMER EAST GERMAN SYSTEM

From the uncertain inception of the explicitly socialist state in the eastern sector, its leadership tried to invest all its inhabitants with a common concern for the unity and progress of the country. The principal vehicle for this was the political federation called the National Front of the German Democratic Republic. In addition to the controlling Socialist Unity Party of Germany (Sozialistische Einheitspartei Deutschlands; SED), it included representatives of the tolerated rump bourgeois parties and of the officially sponsored mass organizations—the Confederation of Free German Trade Unions, the Free German Youth, the Democratic Women's League of Germany, and the East German Cultural League. It was the National Front that, under the guidance of the SED, prepared the lists of candidates for election to the various national and local assemblies, collectively known as Peoples' Assemblies. The right to vote was and remains universal from 18 years of age.

The highest national elected body was the Volkskammer ("People's Chamber"), composed of 500 members. During the socialist era, delegates were elected for five years, and meetings were formal and rare. The Volkskammer elected the Council of State, which met under a chairman who was nominated by the SED and who was formally the country's head of state.

Ministers and ministries of the government had few powers of independent decision; their essential role was

to apply decisions of the SED leadership. Ministries covered a much wider span of national life than in capitalist countries, especially in controlling the entire range of the economy.

There were three tiers of local government: *Bezirke* (districts), rural and urban *Kreise* (counties), and *Gemeinden* (communes). At each level, there was an assembly elected from carefully preselected members. At the local level, voluntary groups such as auxiliaries of the People's Police were organized for community service.

(G.H.K./T.H.El./Ed.)

JUSTICE

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

Structure of the court system

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

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

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

Administrative law

Former East Germany. Except at the lowest levels, the former East German judicial system was mainly concerned with defending the socialist state; protection of citizens' rights and arbitration of disputes were to be the primary concerns of the restructured judiciary. Special (often secret) military tribunals that functioned under the communist government were abandoned. The reconciliation of the West and East German legal codes, which had progressively diverged, was a matter of difficulty.

POLITICAL INSTITUTIONS

The electorate. The voting age was reduced from 21 to 18 in 1970. Both the quadrennial general and provincial elections as well as local elections are attended with the greatest interest and involvement on the part of the electorate. The public is kept informed on political issues through intense coverage in the press, television, and radio, and political affairs provide a topic for frequent debate among German citizens; at no time was this more evident than during the tumultuous and contentious months leading up to reunification in October 1990 and during the local, state, and national elections that followed. The pride with which many Germans once asserted that politics held no interest for them is an attitude of the past. Although voting is not compulsory, an extremely high percentage of citizens participate. Since elections in the *Länder* are staggered throughout the life of each Bundestag, they act as weather vanes of public opinion for the incumbent federal government.

Political parties. Germany's political parties, the sheer proliferation of which contributed to the downfall of the Weimar Republic in 1933, have shown an increasing tendency toward consolidation since the early days of the Federal Republic. Smaller parties have either allied themselves to the larger ones, have shrunk into insignificance, or simply have vanished. Reunified Germany has, in effect, only two large parties, neither of which can easily attain a parliamentary majority. Since 1966, all governments of the Federal Republic have had to hold power through coalition with one of the minor parties. The "5 percent rule," whereby a party may send delegates to the Bundestag or the provincial diets only if it wins at least 5 percent of the vote in a given election, is a highly effective instrument for excluding radical parties of whatever stripe and preventing the formation of splinter parties. Dissent is contained within the wings and factions of the parties themselves.

The party of the centre-right—the "bourgeois" party in older European terms—is the Christian Democratic Union (Christlich-Demokratische Union), which headed Bonn governments from 1949 to 1966 and regained power in 1982. It is established in all *Länder* (including those of the former eastern sector) except Bavaria, where the more conservative Christian Social Union (Christlich-Soziale Union) functions as its counterpart.

Party constituencies

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

The Social Democratic Party of Germany (Sozialdemokratische Partei Deutschlands) is the heir to the Marxist parties dating from the 19th century, and its early postwar leadership maintained a rigorous policy of adherence to classic Marxist doctrine, vehement opposition to the communist movement from which it had split in the early 20th century, and rejection of rearmament for West Germany and its integration into the Western military defense system. In 1959, however, the Social Democrats discarded their doctrinaire approach: the call for nationalizing large industries was forsaken in favour of gradualist reform, and appeals to class warfare were abandoned. The Social Democrats were able, thereby, to broaden their base to attract greater segments of the middle class, whether for reasons of ideology or as an alternative to continued Christian Democratic leadership. Current policies call for full implementation of the reunification program, cooperation with new eastern European democracies, social

and industrial reform, support of NATO, and European unity through the European Communities. The Social Democrats are no longer hostile to the free-market system.

Because neither the Christian Democratic and Christian Social coalition nor the Social Democrats can elect enough delegates to form a viable majority in the Bundestag, since 1963, with few exceptions, the balance of power has been in the hands of the Free Democratic Party through coalition with either of the larger parties. The Free Democratic Party (Freie Demokratische Partei) is the successor of older German liberal parties with a free-trade, probusiness, and anticlerical cast. Today the Free Democratic Party benefits as a small but vigorous party that serves as a liberal, bourgeois alternative.

The Greens (die Grünen) is the environmentalist party and is the only successful, completely new party of the postwar era. Formed by mainly young environmentalists, opponents of nuclear power, and pacifists, the Greens successfully broke the 5 percent barrier in 1983 and in 1987 sent more than 40 voting delegates to the Bundestag.

A number of small and fringe parties of special-interest groups, as well as extremists of the left and right, are more visible than effective. The German Communist Party (Deutsche Kommunistische Partei), successor of the historical Communist Party of Germany, outlawed in the Federal Republic in 1956, is without consequence. The extreme right, exploiting the themes of nationalism and xenophobia, is mainly represented by the National Democratic Party of Germany (Nationaldemokratische Partei Deutschlands), the German People's Union (Deutsche Volksunion), and the Republicans (Die Republikaner), who managed in 1989 to penetrate the 5 percent barrier in a few local elections, without, however, winning seats in the Bundestag.

From one party to many

Communism and its collapse. In the former Democratic Republic, the Socialist Unity Party (SED) was the product of the enforced merger in 1946 of the Social Democratic Party with the Communist Party. The two bodies that actually wielded power were the Secretariat of the Central Committee of the SED and the Politburo. Both bodies were dominated by the aging party leadership, dating from the Stalinist period. Both were chaired by the first secretary of the Central Committee and leader of the SED; between 1971 and 1989 this position was held by Erich Honecker. The SED was all-pervasive; in addition to a hierarchy of local and regional branches, there was at all levels of the economy and in the government, the universities, and other institutions a parallel SED structure to control the implementation of party decisions. There were similar parallel structures in the trade unions and the Free German Youth, leading to a massive bureaucratic proliferation.

The precipitous events of 1989 toppled the SED and opened the way for the first free elections in eastern Germany since before the Nazi dictatorship. Numerous parties emerged, and March 1990 elections produced a coalition government led by the conservative Christian Democratic Union and including the Social Democratic Party, the German Social Union, and Democratic Awakening, as well as a liberal alliance supported by West Germany's Free Democratic Party. The Party of Democratic Socialism, as the discredited SED now called itself, received no share in the coalition. Ironically, the reform movement that had catalyzed the country's democratic revolution failed to constitute a party in its own right. In the local and state elections held after reunification in October, the Christian Democratic Union, led by Chancellor Helmut Kohl, won a sound vote of confidence.

ARMED FORCES

The Federal Republic has been a member of the North Atlantic Treaty Organization (NATO) since the implementation of the Paris Treaties in May 1955. Throughout its existence, West Germany was the only NATO country of western Europe with territory bordering two member nations of the Warsaw Pact, and planning from the early days of the Western defense system was posited on its strategic geographic location, which made West Germany vulnerable to an armed invasion and the possible site of a land war. A land strip barely 60 miles wide separated its border with the Warsaw Pact nations from a third of its population and a fourth of its industry. All successive Bonn governments were unwavering in their commitment and support to the aims of NATO, and the Federal Republic's continued membership in the organization was a central tenet of the 1990 reunification agreement.

Branches of service. The German contribution to the Western defense system, apart from playing host and contributing to the continued presence of NATO troops on its soil, takes the form of its combined arm of defense known as the Federal Armed Forces. Forming the largest contingent of NATO troops in Europe, the armed forces are divided into an army, navy, and air force. From its inception it was envisioned as a "citizens' " defense force, decisively under civilian control through the Bundestag, and its officers and soldiers are trained to be mindful of the role of the military in a democracy. Conscription for males is universal, the military liability beginning at 18 and ending at 35 years of age. The period of compulsory active service is 18 months. The right to refuse to enter military service on grounds of conscience is guaranteed under Article 4 of the Basic Law; instead, a recognized conscientious objector must perform 20 months of socially useful service.

The Federal Armed Forces

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

The former eastern sector. Like the Federal Republic, the Democratic Republic had compulsory military service for men. Equipment and training were modeled on those of the Soviet Union, which still had some 380,000 men in the country as late as 1990. Military morale and organization suffered greatly in the events of late 1989. The special Border Troops set up to guard the Berlin Wall and the boundary with West Germany became unnecessary, and the factory militia, once used to support communist power, was disarmed and dissolved.

A treaty signed by the Federal Republic and the Soviet Union in July 1990, in preparation for German reunification, provided that the two countries would arrange the withdrawal of Soviet troops from former East German territory over a period of three to four years and that Germany would help to finance this transition. Correspondingly, no foreign NATO troops would be stationed in the territory, with the exception of troops of other Allied powers stationed in Berlin during the interim. International, as well as domestic, concerns regarding the potential military strength of a united Germany were addressed by a government declaration limiting the country's total ground, air, and naval forces after four years to no more than 370,000 troops.

Soviet withdrawal

LAW ENFORCEMENT

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

Former East Germany. In the Democratic Republic, the hated State Security Police, or "Stasi," which was responsible for the suppression of dissent, disappeared with the disintegration of the Honecker government in late 1989. Under the freely elected interim government prior to reunification, the German People's Police, the regular police force, continued on a strictly civil basis, operating under the law.

EDUCATION

Preschool, elementary, and secondary. Schooling is free and compulsory for all Germans from their 6th through 18th year of age. At age 15 or 16, part-time classroom attendance is combined with vocational training for most young people. While control of education is the sovereign prerogative of the *Länder,* a permanent commission strives for a certain uniformity in curriculum, requirements, and standards, the implementation of which may vary, depending on the priorities of the individual *Land.* Books and study materials are usually free, and financial assistance and other forms of support are available in cases of hardship.

Although many reforms have been introduced and alternative plans are available, the traditional pattern still prevails. A common elementary school is attended for the first four years. A determination is then made as to whether a pupil will attend a further five to six years of elementary school and then, at 15 or 16 years of age, be apprenticed to a trade or trained in a special industrial vocation; will attend a school essentially preparing for clerical, administrative, or commercial careers; or will receive lengthy academic training in preparation for entry into university or the higher ranks of business or the civil service.

Elementary and secondary education

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

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

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

There are now a small number of so-called *Gesamtschulen* (equivalent to British comprehensive schools). These are intended as alternatives to the previously rigid division into three levels, often criticized for forcing the choice of a child's future at too early an age, a choice that, once entered upon, was almost impossible to change.

The postwar increase in the number of pupils attending *Gymnasium* has created a crisis for German universities.

The German university system

The ancient tradition of *Lehrfreiheit* ("academic freedom") that went hand in hand with *Lernfreiheit* (*i.e.,* the freedom of students to pursue studies however they saw fit and for however long it took) has had to yield place to the *numerus clausus* ("closed number") system similar to the tradition in Anglo-American universities of a restricted number of places.

Higher education. The German universities, famed in history and noted for their enormous contributions to learning, especially in the 19th and early 20th centuries, have been put under severe strain by the swelling numbers of students and changing social conditions. It has become all but impossible for students to take as long as they wish to complete their studies or to move from university to university as they please. Lecture rooms, seminars, and libraries are disastrously overburdened, and, as in other countries, the explosion in knowledge and in higher education has undermined the usefulness of original research. In addition, the solitary, authoritarian position of the one professor who stood at the pinnacle of each discipline in each university has been weakened.

To meet the demand for higher education, the Federal Republic has steadily added to the number of existing universities. It has created entirely new academic universities to add to the ranks of the ancient universities, and it has upgraded the status of institutes and colleges of technology, education, and art to university rank, while creating new specialized or technical institutions such as the *Fachhochschule,* a higher technical college specializing in a single discipline, such as engineering, architecture, design, art, agriculture, or business administration. Little or no prestige or social distinction is attached to whether a student has studied at Heidelberg, founded in 1386, or at one of the 20 or more universities established after 1960. One innovative university is that found in Hagen, where the teaching is largely accomplished by correspondence through regional study centres. Whereas in East Germany research institutions were generally separate from universities, the Western system combining research and teaching was implemented nationally following reunification.

There are more than 90 universities or institutions of equivalent rank. The doctorate is the only degree offered, although the Magister, or Master of Arts, abandoned in the 17th century, has been partially revived.

There is an extensive range of possibilities for extended education or extramural studies. Upward of 1,500 *Volkshochschulen* ("people's universities") enroll some two million adults for complete courses or individual subjects, whether in preparation for, or furtherance of, a career or out of personal interest. The government has also promoted the retraining and further vocational education of workers.

The former East German system. Education in East Germany reflected the values of the communist state. Because of the expectation that all women would work, except in the period of childbirth, there was a full provision of day nursery (*Kinderkrippe*) facilities for children up to 3 years of age, kindergartens for those aged 3 to 6 years, and after-school centres (*Horte*) for those aged 6 to 10 years. The unified educational system applied nationally determined syllabi and in principle gave equal opportunity to all; no fees were payable. The core of the system was the public polytechnic school, *Oberschule,* which gave a uniform general education to virtually all children between the ages of 6 and 16. From grade seven, pupils received actual work experience in a factory or on a farm. There were special classes and schools for pupils gifted in such fields as mathematics, modern languages, music, and sports, as well as provisions for the handicapped. At grade 10, selection took place for entry into the two-year *Erweiterte Oberschule,* which led to the high school diploma (*Abitur*) and ensured entry into higher education. Because selection was designed to produce only the number of graduates the country needed, competition at this stage was intense, being based not only on scholastic achievement but also on social contributions (Free German Youth) and parental status (children of workers being preferred). Most pupils continued through another two years of vocational training, with preference given to "new" trades such as microelectronics or computer science; after a year's work experience, some of these students moved to technical and engineering colleges (*Fachschulen*).

The *Oberschule*

As in other fields, the events of October 1989 introduced confusion and change into the educational system. Academics united to reject dictation by communist party branches and to question the validity of appointments that had been based on political rather than academic considerations. Students rejected the Free German Youth, which had done so much to rule their lives, in favour of nonpolitical associations, and compulsory military training in schools was abolished. Compulsory Marxism gave way to a more balanced philosophy. The syllabus of social studies classes, previously dominated by the achievements of the

communist government, was returned to the discretion of the teachers.

HEALTH AND WELFARE

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

Insurance and services. Health and retirement insurance are compulsory for all workers and employees earning below a certain level of income geared to the cost of living. Under German labour law, a categorical distinction is made between workers and clerical and managerial employees, and differing rules and rates apply to each group. Employees above a certain salary level or self-employed persons are generally exempt from most obligatory payment systems; however, although the former usually participate in a firm's retirement plan, almost all persons in the upper salary brackets or the self-employed are covered by private insurance as comprehensive as the government-sponsored plans. With increasing prosperity, an increasing percentage of the population is subscribing to these somewhat more expensive but also more generous private plans. Almost all of the population is covered by compulsory health insurance, and the Federal Republic ranks highest among comparable nations of continental western Europe in the proportion of money allowed toward health costs—about 90 percent of all medical costs incurred. Contributions range from about 8 to 13 percent of wages or salaries.

Compulsory national and private insurance plans

Medical care is of a high standard, and rural areas are well served. Hospitals are usually operated by municipalities or religious bodies or as proprietary operations owned by one or more physicians. Public health standards are high. A triumph of the public health system has been the conquering of tuberculosis, a disease formerly endemic to Germany but now rarely encountered.

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

The former East German system. The German Democratic Republic developed an extensive state health-care system. Financial support and free medical assistance, medicine, and nursing care were made available in case of sickness or accident. The scheme, available to all citizens, was funded in part by a uniform and obligatory state insurance system, managed for most people by the trade unions. There was a pension system for the aged. In late 1989 the health system was devastated by the migration of doctors and nurses to West Germany. The insurance system was then removed from the East German trade unions and aligned on the various employment, health, and retirement insurance systems of the Federal Republic.

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

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

East German payment

HOUSING

Much of early postwar housing construction was aimed at the immediate relief of the acute housing shortage caused by wartime destruction. The emphasis later shifted to quality construction and to satisfying a higher standard of living, from built-in bathing and sanitary facilities to a larger area of living space per person. The average area per dwelling unit increased from about 750 square feet in 1960 to 970 square feet in 1986, an average living space per person of 370 square feet.

The demand for housing is not completely in balance with the supply. There is a shortage of dwellings in the medium range of price and size, while there is a glut of luxury units selling at high prices. Prewar housing, much of it of solid and capacious construction, has undergone renewal, and for many Germans, living in a refurbished building in the older districts of the large cities is preferable to occupying newly built structures.

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

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

"Social housing"

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

The former eastern sector. In principle, under the communist government of the Democratic Republic each citizen and his or her family had the right to adequate accommodations. Initial investment, however, was limited by the direction of resources toward industrial reconstruction. Only after the 1970s was there a major program to develop high-rise housing projects. Rents everywhere, together with charges for heating and electricity, were held at extremely low, subsidized levels. After 1989 it was proposed that rents should rise in steps to realistic market levels, in line with wage increases to West German levels in preparation for reunification. The massive migration of East Germans and other refugees into West Germany in 1989 severely strained the urban housing market. At reunification, the problem remained closely tied to that of employment opportunity.

Housing policy

STANDARDS OF LIVING

Postwar prosperity. The standard of living in the Federal Republic is one of the highest in the world. The distribution of wealth compares favourably with that of other industrial nations. Powerful incentives to save are offered by the state not only in the form of housing subsidies and tax concessions but also through bonus saving schemes. Thus, for persons whose income does not exceed

a certain level, savings of up to a fixed amount kept in a bank or savings institution for six or seven years will be granted a bonus of 14 percent of the principal by the government. Payments into a longer-term savings agreement, a home-savings contract, or a similar instrument are given a "worker savings grant" by the state.

Earning power for both workers and employees assures an adequate income to meet the cost of living. There is no exaggerated difference between the earning power of blue-collar workers and white-collar employees, salaries tending to be only moderately higher than wages. While above a certain level of management incomes and benefits rise precipitously, chief executive officers in Germany earn only 6 to 7 times the average worker's pay (as opposed to more than 13 times in the United States and more than 9 in the United Kingdom). Since, as in all the European Communities countries, a major portion of tax revenue derives from excise levies and the value-added tax, low- and medium-income workers collectively bear the greater tax burden.

Life under communist rule. Life in the Democratic Republic reflected the German traditions of order, discipline, respect for authority, and work. On the other hand, Soviet influence was apparent, not only in a pattern of harsh control at all levels of society but also in a divergent vocabulary (*Politburo, Kombinat, Dacha*) and even in diet. Following the Soviet example, production was central; around it all other activities revolved. This was reflected in the position of women, whose full-time participation in the labour force, however reluctant, was assumed. Backed by adequate child-care facilities, women could and did enter all but the most dangerous and arduous careers, but they were heavily concentrated in such fields as clerical work, retailing, teaching, and medicine; even so, the top positions tended to be held by men, as were the higher positions in politics (perhaps reflecting the fact that formal equal opportunity for women did not compensate for interruptions for childbearing).

Beginning with nurseries, the educational system socialized children and young people in responsibility for the needs of others and of the community. Important vehicles in this process were the Pioneer organization for children 6 to 14 years of age and the Free German Youth thereafter. Heavily subsidized youth and sports clubs offered welcome opportunities for teenagers to meet outside their crowded homes, and neighbourhood restaurants were reserved at least one evening a week for youth dances.

Generally speaking, East Germans became accustomed to a life in which guaranteed employment removed the dangers of competition and in which most of the requirements of life were provided free or at low cost by the state. Yet there was much resentment of the "nanny state," perhaps stimulated by radio and television broadcasts from the West. Collapse of the communist regime gave rise to a majority mandate for German reunification. Yet there were some who feared being overwhelmed by Western capitalism and an accompanying loss of the tradition of community care that had been one of the positive features of the Democratic Republic.

Cultural life

THE CULTURAL MILIEU

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

The cultural life of the Democratic Republic reflected the rich German cultural heritage but was powerfully molded by a socialist ideology of predominantly Soviet inspiration. Artistic works were expected to be realistic and accessible to the people. Guidance in the required direction was provided by exhortation through a range of associations and by some degree of censorship; the state, as virtually the sole market for artistic products, inevitably had the last word.

Despite the decades of political division, the German cultural and artistic tradition remains unitary, and it is shared with three other countries, Austria, Switzerland, and Liechtenstein. In the matter of art and literature, the adjective *deutsch* ("German") still knows no strict political boundaries. The Austrian Gustav Mahler is, for example, called a "German" composer and the Swiss Friedrich Dürrenmatt is a "German" playwright because they are in the German cultural tradition.

STATE OF THE ARTS

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

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

Theatrical and musical traditions

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

Traditional arts and crafts. The incursions of modern patterns of life have done much to weaken the traditional arts, entertainments, and customs of regional and rural Germany. In southern Germany, however, the older arts and usages have persisted concurrently with a gradual adaptation to a modern, urban pattern of life; the old and the new coexist in an incongruous compatibility. The young still dance around the village maypole, but they also dance to the disco beat. The woodcarvers, violin makers, and gunsmiths of Upper Bavaria continue, under great economic pressure, to follow their trades, not because it is quaint to do so but because they still believe in the work itself; the rural women in the Black Forest still wear elaborate costumes known as *Tracht* on festival days, not to amaze tourists but because they have always done so—yet these are the areas in which the tourist industry is most highly developed. Some usages have all but disappeared in

the villages: older women now seldom wear black dresses and scarves, and the village men no longer appear in top hat and cutaway for a funeral procession.

Popular festivals

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

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

The German theatre has long been faced with the dilemma of either relying on the rich repertoire of German classics from the 18th and 19th centuries, along with a restricted number of 20th-century dramas and contemporary plays in translation from British, American, and French authors, or of being bold and innovative. Small experimental theatres enjoy a lively and hazardous existence in the major cities and often closely resemble Germany's long-lived and still lively convention of political cabaret.

In the Democratic Republic, all theatres were owned by the state. The Deutsches Theater in Berlin reopened in September 1945 and was the first German theatre to perform following the Nazi collapse. The old German National Theatre in Weimar was the first to be rebuilt after 1945. Understandably, Berlin dominated the development of theatrical ideas, especially through the work there of Brecht (1884–1977) and his successors at the Theater Am Schiffbauerdamm.

Bertolt Brecht in Berlin

Music and dance. German concert halls, like the theatres, face the dilemma of satisfying the public's preference for 18th-, 19th-, and early 20th-century composers or of giving contemporary works a better hearing. The works of Hans Werner Henze, whose opera *The Young Lord* is widely performed abroad, the symphonic works of Gottfried von Einem, and the electronic music of Karlheinz Stockhausen have established these men as the most widely known of contemporary German composers. The Berlin Philharmonic, the Munich Philharmonic, the Bamberg Symphonic Orchestra, and the Stuttgart Chamber Orchestra rank with the leading world ensembles. Innumerable German voices of opera and concert have secured a place in the vocal pantheon—Elisabeth Schwartzkopf, Dietrich Fischer-Dieskau, Christa Ludwig, Brigitte Fassbaender, Peter Hofmann, René Kollo, Wolfgang Brendel—and the opera houses of Hamburg (the oldest), Berlin, Cologne, Frankfurt, and Munich rank with the world's greatest. The system of state-supported opera has allowed many young North American and British voices to train in Germany, some remaining in the permanent companies there.

Major orchestras

East Germany attained an international reputation for musical excellence, as exemplified by the Leipzig Gewandhaus Orchestra under the direction of Kurt Masur, the choir of St. Thomas Church in Leipzig, and individual artists such as tenor Peter Schreier. The state contributed by the restoration of buildings, such as the Berlin Opera House on the Unter den Linden and the Berlin Schauspielhaus, a theatre associated with Schiller that was converted into a concert hall. New buildings included the opera house and concert hall for the Gewandhaus orchestra in Leipzig.

A conspicuous international success in the performing arts has been that of the Stuttgart Ballet at the Württemberg State Theatre; it achieved world prominence under the South African-born John Cranko, and its success was continued after his death in 1973 under Marcia Haydée. New forms of dance have been pioneered in Wuppertal as *Tanztheater* ("theatre of the dance") by the choreographer Pina Bausch. The Hamburg Ballet is a lively centre of world ballet.

The visual arts and architecture. Between Expressionism, the characteristic German art movement of the early 20th century, and the decades of the postwar period lay a void that came to be occupied by latter-day surrealists, such as Edgar Ende, Mac Zimmermann, and Richard Oelze. Leading graphic artists include Horst Janssen, Anselm Kiefer, and Hans Bellmer. The tradition of the older Expressionism has been resumed with such groups as the *Neue Wilde* ("New Wild Ones") working in Cologne, Berlin, and Hamburg, and such artists as Elvira Bach and Jörg Immendorff have been acclaimed abroad. The latter is also noted as a sculptor, along with Karl Hartung and Rolf Szymanski in Berlin, Erich Hauser and Otto Herbert Hajek in the south, and Ulrich Rückriem, Norbert Kricke, the late Joseph Beuys, and others in the Rhineland.

Under the communists, much art was openly didactic or propagandist. Sculpture could be literally monstrous, as in the towering 148-foot (45-metre) Lenin monument in East Berlin (admittedly by a Soviet sculptor), but there were examples at a more human scale, such as the layout of the Marx-Engels-Platz in the same city.

German architecture

German architecture today—indeed world architecture today—is very much the creature of the Bauhaus school that was associated with Weimar and Dessau in the 1920s and that evokes the names of such giants as Walter Gropius and Ludwig Mies van der Rohe. But its dogmas—strict harmony of style with function, insistence on the intrinsic beauty of materials, and puritan disdain of decorativeness—dutifully mirrored in city after city, building after building, in the postwar years began to yield in the 1970s and '80s in the Federal Republic as elsewhere to the more free-ranging Postmodern style. Some leading German exponents of this newer school, which has sought to reverse the unremitting rectangularity of the International style, have been Josef Paul Kleihues, Oswald Mathias Ungers, and two Austrians, the brothers Robert and Léon Krier.

In the early postwar years in the Democratic Republic, architecture paralleled the other arts in reflecting socialist ideological influence modeled on Soviet designs; several major showpieces were carried out, but sometimes exaggerated decorations and eclectic designs led to tasteless results, as, for example, the overall appearance of the Karl-Marx-Allee in East Berlin. In an attempt to remedy the situation, principles of urban construction closer to contemporary Western forms were laid down to serve as the foundation for ensuing architectural design, although the effect was rather monotonous. Except where ideological factors intruded (destruction of the Berlin Palace), the Democratic Republic had an admirable record for the preservation of ancient buildings.

Film. In contrast to the situation after World War I, when Germany helped set the pace in many newer forms of art and entertainment, most notably in the genre of motion pictures, the two Germanys were slow in finding métiers of their own in the cinema in the early decades after World War II. The likes of Fritz Lang, Ernst Lubitsch, F.W. Murnau, and G.W. Pabst, who had virtually defined the cinematic art in the 1920s and early '30s, had no counterparts to rescue the German film from the slough of mediocrity into which it had fallen as a result of the Third Reich. But in the late 1960s a new generation of young German film producers and directors began to reestablish West Germany as a serious filmmaking country. Alexander Kluge, Volcker Schlöndorff, Werner Herzog, Wim Wenders, Hans-Jürgen Syberberg, and Wolfgang Petersen have

matched foreign critical acclaim with commercial success. The prodigious work of Rainer Werner Fassbinder, whose following in and out of Germany amounted to a cult, was cut short by his early death in 1982. Women directors have earned special recognition in the contemporary German cinema, including Margarethe von Trotta, Marianne Rosenbaum, and Helma Sanders-Brahms.

Arts festivals. The arts are cultivated and commemorated in Germany on a scale scarcely equaled in any other country, with a proliferation of festivals. Most of the major cities and scores of small towns or villages sponsor festivals that celebrate all genres of music, film, and the performing arts. Among the most renowned of these are

The Bayreuth Festival

the Bayreuth Festival, the celebration of the music of Richard Wagner held each summer in northern Bavaria and still under the direction of Wagner's descendants. The oldest festival is the Passion Play held every 10 years since 1634 in Oberammergau in southern Bavaria to celebrate the town's deliverance from the plague.

Berlin alone has five major festivals: the Berliner Festwochen ("Berlin Festival Weeks") in September and October, with five weeks of musical events; the JazzFest Berlin in November; the International Film Festival in February and March; Theatertreffen Berlin ("Berlin Theatre Meeting"), featuring plays from throughout the German-speaking world; and Horizonte ("Horizons"), a festival of world cultures. Munich has an opera festival in July and August, with emphasis on Richard Strauss. Festivals in Würzburg and Augsburg are dedicated to Mozart; while Ansbach has a Bach festival and Bonn a Beethoven festival. Others of note include the International May Festival in Wiesbaden and the Festival of Contemporary Music in Donaueschingen. The greatest event in German music, the Salzburg Festival, is held across the Bavarian border in Austria.

Publishing. Germany is a major publishing nation, its production being surpassed only by the United States and the Soviet Union. Germany traditionally has been the land of the small- and medium-sized publishing house. The Bertelsmann group, however, is Europe's largest publisher. In what was formerly the western sector, book publishing is centred in no one city but is concentrated fairly evenly in Berlin, Hamburg, and the regional metropolises of Cologne, Frankfurt, Stuttgart, and Munich. Leipzig was the publishing capital of the Democratic Republic and remains an important centre of the industry. Gotha is renowned for the production of maps and atlases.

The press. Germany is very much a land of newspaper and periodical readers. Freedom of the press is guaranteed under the Basic Law, and the economic state of its several hundred newspapers and thousands of periodicals is enviably healthy. Most major cities support two or more daily newspapers, in addition to community periodicals, and few towns of any size are without their own daily.

The press is free of government control, and no newspaper is owned by a political party. In 1968 a government commission set out restrictions on the total circulation of newspapers or magazines that could be controlled by one publisher or group. While newspaper and periodical ownership thus cannot be the monopoly of any one ownership, one-fifth of newspaper circulation is in the hands of the Axel Springer AG conglomerate, which also publishes magazines and books. Other major newspaper proprietorships are those of the Westdeutsche Allgemeine Zeitung, the Süddeutscher Verlag, the Verlag DuMont Schauberg, and Societäts-Verlag. A German Press Council founded in 1956 sets out guidelines and investigates complaints against the press.

A national press exists on one level in the form of *Die Welt* (Hamburg and Bonn) and the *Frankfurter Allgemeine Zeitung,* together with the *Stuttgarter Zeitung* and the *Süddeutsche Zeitung* (Munich) and the *Frankfurter Rundschau,* which also command special international circulation and respect. Another level of the national press is represented by the universally circulated tabloid *Bild-Zeitung* (Hamburg), with eight regional editions.

The Sunday press

The major editions of newspapers in the former western sector, with additional features on politics and the arts, appear on Saturday. A lively Sunday press complements the daily newspapers; *Die Zeit* (Hamburg) is the most prestigious and influential example. Others include the venerable *Rheinischer Merkur* (Coblenz), the *Bayernkurier* (Munich), and the *Deutsches Allgemeines Sonntagsblatt* (Hamburg). Sunday counterparts of the major dailies, *Welt am Sonntag* and *Bild am Sonntag,* are run virtually as separate newspapers.

The genre of the *Illustrierte* dominates the German magazine market. Some of these popular weekly glossies, such as *Stern, Quick,* and *Bunte,* carry features, including investigative reporting, of a high calibre; others, however, cater to an unquenchable public thirst for the escapades of celebrities, bizarre crime, the annals of gracious living, and sundry escapist topics. Apart from a wealth of specialized journals and quality business-oriented magazines, the role of high-prestige magazines of opinion is largely subsumed by the weighty weekend editions of the quality press.

A special niche is occupied by the weekly newsmagazine, *Der Spiegel,* a journalistic power in its own right, which, since its founding in the period immediately after World War II, has shaped public opinion in the Federal Republic through its editorial posture of the skeptical, nonaligned observer and guardian of the public conscience. Exhaustive in its coverage and polemical in tone, it features thorough, critical investigations of events from both the past and the present.

Throughout the socialist era in the Democratic Republic, the press (as well as film, radio, and television) was firmly in the hands of the state. After the events of 1989, independent newspapers began to appear, and even the Party's *Neues Deutschland* exhibited a livelier content than the verbatim reporting of speeches by communist statesmen. Freedom of the press brought a tide of new publications.

Broadcasting. German radio and television are not state-controlled. Only public corporations, however, were permitted to broadcast until the mid-1980s, when a dual system of public and commercial stations came into effect. A ruling of the Federal Constitutional Court in 1986 held that the public corporations comprised the "basic supply" of news and entertainment and commercial outlets a "supplementary supply." Licensing and control of public broadcasting is under the Federal Ministry of Post and Telecommunications. Support of public broadcasting comes from fees paid by the owners of radios and television sets.

The public corporations enjoy great freedom in establishing their own broadcasting policies. Attempts to control these policies, which are often hostile to incumbent governments, have been repeatedly rebuffed, so that, in practice, German television, more than radio, enjoys remarkable latitude and independence. During the partition, the state control of broadcasting, customary for all Soviet-style countries, was modified for East Germany in that television programs from West Germany and West Berlin could be received throughout most of the country.

Public radio and television are arranged along regional lines, with nine corporations offering two to four radio programs and combining to form one evening television program, Channel One, or ARD (Arbeitsgemeinschaft der öffentlich-rechtlichen Rundfunkanstalten Deutschlands). This is complemented by a second television network, Channel Two or ZDF (Zweites Deutsches Fernsehen), based in Mainz. A third program is operated by ARD but is organized and broadcasts along regional lines, with special emphasis on local and regional events and school instruction. The uneven quality of entertainment in both radio and television is offset by the high-quality news coverage and political and social reporting that go to make the German public one of the best-informed of any nation.

Recreation. In a mere generation or two, leisure has become a major pursuit, if not an industry, in German life. As did workers in most Western industrial countries, Germans once toiled under an unrestricted six-day work week, often 10 hours a day, with breaks only on Sundays and for major feasts and festivals. Added to the feasts and festivals now are several secular holidays, a work week of 40 hours or less, and three weeks or more of paid vacation time. The result is that Germans have more leisure time than workers of any Western country. What to do with

this abundance of leisure has become a national preoccupation, and there is even a German Leisure Association that conducts research on leisure activity and dispenses information and advice.

A duality of traditional and modern characterizes the German pattern of recreation. To a surprising extent in a country so highly industrialized, ancient feasts and usages, some antedating Christianity, are still observed on a wide scale in both Roman Catholic and Protestant areas, especially pre-Lenten celebrations known as *Fasching* in the southern regions or *Karneval* in the Rhineland. Added to the major religious festivals are a wealth of local celebrations—wine festivals, beer festivals, harvest festivals, hunting festivals, historical festivals—the so-called *Volksfeste* that are deeply rooted in German custom, the most famous example of which is Munich's Oktoberfest, held each September.

The German tourist

Travel has become the favoured pastime of a majority of Germans. More than half of all adults in the Federal Republic take at least one annual trip for pleasure, and many take more than one. Many Germans now take a winter vacation and a summer vacation. Older persons often visit one of the multitude of spas to rest and recuperate, in addition to a holiday trip for pleasure. More Germans take trips abroad for pleasure than the citizens of any other country.

Under communist rule in East Germany, sports were a matter of national significance. Young athletes of promise were chosen for special training and education. The Leipzig College of Physical Culture trained the athletic elite, with a view toward securing prestige in international competition. The college's associated medical school developed methods, including the use of drugs, to enhance athletic performance, particularly by women. After the events of 1989, complaints emerged regarding the lack of sports provision for ordinary members of the population.

CULTURAL INSTITUTIONS

International cultural activities

Great importance has been attached in the Federal Republic to the support of the nation's cultural, educational, and scientific resources through institutions supported in whole or part by public funds. A prodigious complex of organizations is devoted to acquainting the international public with the culture, life, and language of the German peoples and familiarizing the German public with the culture and life of other nations. Cultural representation abroad and cultural exchanges are maintained in abundance with the advanced industrial nations of the West and, increasingly, with eastern Europe, but special emphasis is laid upon the development of educational and cultural ties with the developing nations of the world. Germany not only has assumed a major role in lending reserves of technological skill and capital to developing the resources of the nonindustrial or poorer nations, but it also has become a major centre for the education and training of students from these countries in the professions, the sciences, and technology.

Prominent among cultural groups is the Goethe Institute of Munich. Founded in 1951, it has some 150 branches around the world. It operates schools in the Federal Republic and abroad to offer instruction in the German language. It maintains lending libraries and audio-visual centres, sponsors exhibits, film programs, and musical and theatrical events, and often brings personalities of note in the arts to visit its branches, including the most prominent overseas centre, the Goethe House in New York City.

Museums and galleries. Germany has thousands of museums of all descriptions, from those housing some of the world's great collections of fine art, archaeological relics, or scientific displays to those with exhibitions of minutiae, such as the playing-card museum in Stuttgart. Some of the museums and galleries of great note are the Alte Pinakothek ("Old Pinacotheca," or "Art Gallery"), the Neue ("New") Pinakothek, and the German Museum in Munich; the Pergamon Museum and the museums of the Prussian Cultural Property Foundation in Berlin; the Germanic National Museum in Nürnberg; the Roman-Germanic Central Museum in Mainz; the Senckenberg Museum of natural science in Frankfurt am Main; and the State Gallery in Stuttgart. Some museums are highly specialized, devoted to a special artist or school or genre, but many combine natural science and fine arts, as in Hannover and Wiesbaden. There is a large number of ethnological museums, such as the Linden Museum in Stuttgart and the Museum of Ethnology in Berlin's Dahlem Museum. Important art treasures are scattered in the scores of smaller museums, libraries and archives, castles, cathedrals, churches, and monasteries throughout the country. Specialized exhibits that gather all the available art, artifacts, or documents pertinent to a given theme attract visitors from throughout the world—for example, the 500th anniversary celebration of Albrecht Dürer in Nürnberg in 1971, the Hohenstaufen exhibit in Stuttgart in 1977, or the exhibit in West Berlin in 1981 commemorating Prussia.

West Germany's libraries

Libraries. Among the nation's great libraries are the Bavarian State Library in Munich (the largest) and the library of the Prussian Cultural Property Foundation in Berlin (formerly the Prussian State Library). The German Library at Frankfurt am Main is the country's library of deposit and bibliographical centre. The Technical Library at Hannover is the Federal Republic's most important for science and technology and for translations of works in the fields of science and engineering. Apart from the great university libraries at Heidelberg, Cologne, Göttingen, Tübingen, and Munich, a wealth of ancient manuscripts, early printed works, and manuscripts and documents from the Renaissance to the present are dispersed in smaller collections. The great research libraries are complemented by an extensive system of lending libraries operated by the *Länder,* the municipalities, the library associations of the Roman Catholic and Evangelical churches, and other public associations and institutes. Virtually no citizen is out of easy reach of a library. The German State Library in Berlin and the German Library in Leipzig served this same function for the Democratic Republic, the latter institution including a collection of almost all the literary output in German since 1913. Libraries throughout the country benefited from the mutual accessibility of formerly divided resources made possible by German reunification.

(G.H.K./H.-J.K./R.E.H.M./T.H.El./Ed.)

For statistical data on the land and people of Germany, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

HISTORY

Ancient history

The Germanic consonant shift

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

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

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

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

Social structure

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

COEXISTENCE WITH ROME TO AD 350

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

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

Transformation of the Germanic world

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

THE MIGRATION PERIOD

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

The Roman Empire nevertheless remained an important power in Europe, having both troops and funds. Hence Germans on the run from the Huns were anxious to make peace; even the Visigoths accepted a settlement in Gaul in 418. Since Germanic groups had no sense of Germanic nationalism, they could be played off against one another; thus Vandals were savaged by the Visigoths between 416 and 418. Until about 450 fear of the Huns meant that the empire could, in moments of crisis, mobilize at least Visigoths, Burgundians (received into Gaul after being defeated by the Huns in 439), and Franks for its defense. Soon after Attila's death in 453, however, the Hun empire collapsed, and Rome lost this diplomatic weapon. It also suffered a progressive loss of revenue as territories were either occupied, or—like Britain—declared themselves independent.

The rise of independent kingdoms

The balance thus swung further in the Germans' favour, and they eventually declared themselves independent. In the 470s a Visigothic kingdom emerged in southwestern Gaul and a Burgundian kingdom in southeastern Gaul, and Clovis created a Frankish kingdom in the north. The Vandals already controlled North Africa and the Suebi part of Spain. Ostrogoths freed by the collapse of the Hun empire conquered Italy between 489 and 493, and Gepidae and Lombard kingdoms dominated the Danube.

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

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

Merovingians and Carolingians

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

MEROVINGIAN GERMANY

The beginnings of Frankish domination

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

THE RISE OF THE CAROLINGIANS AND BONIFACE

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

Creation of the German church

CHARLEMAGNE

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

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

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

The subjugation of Saxony

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

THE EMERGENCE OF GERMANY

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

Germany and France

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

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

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

Changes in social organization

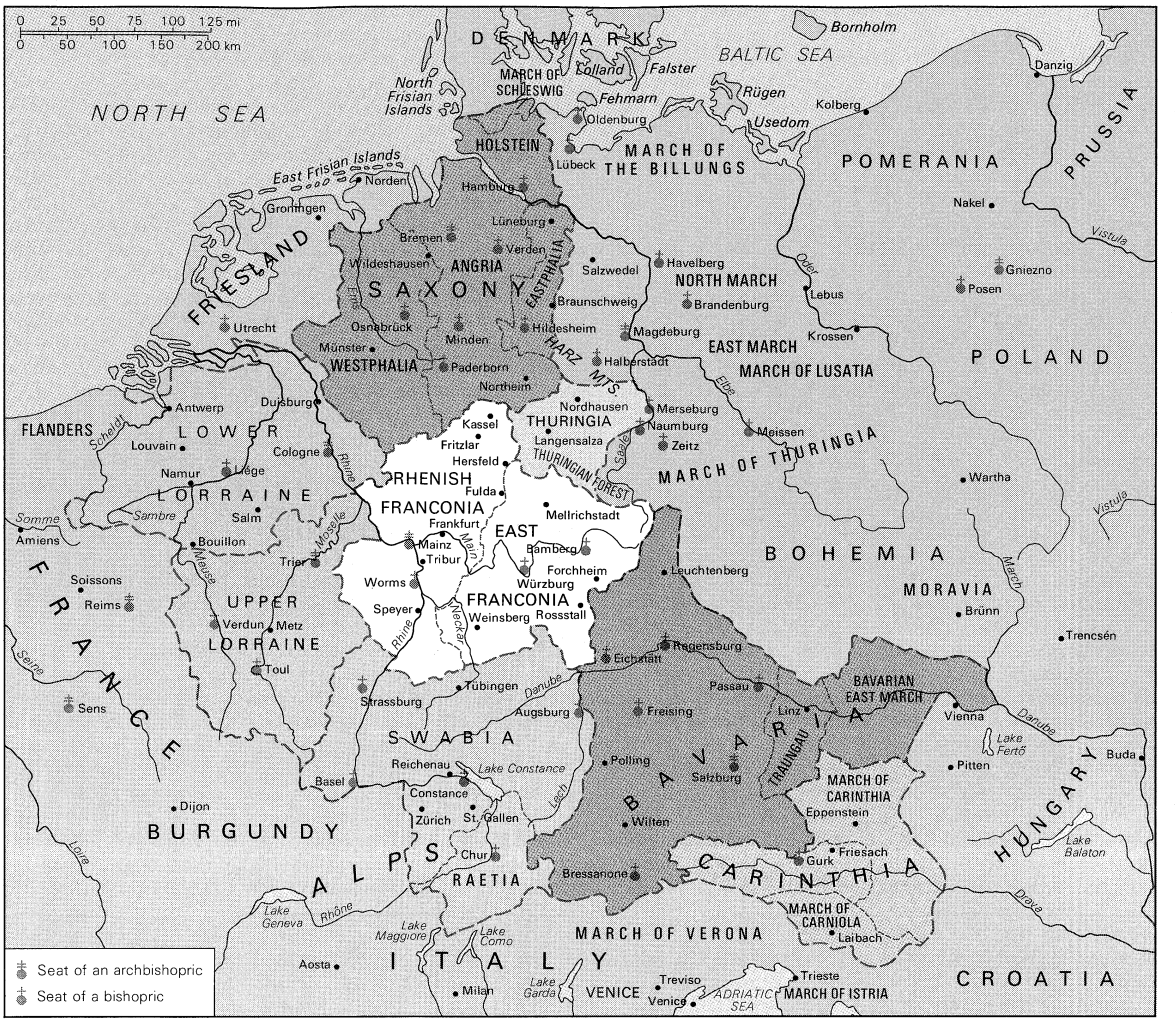
Germany from 911 to 1250

THE 10TH AND 11TH CENTURIES

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

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

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



Germany in the 10th and 11th centuries.

Development of serfdom

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

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

The election of Henry I the Fowler

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

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

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

Reliance on the clergy

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

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

Defeat of the Hungarians at the Lech

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

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

Legacy of savagery

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

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

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

The rise of the advocates

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

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

Power of the kings

The king was the personal overlord of all the great. His court was the seat of government, and it went with him on his long journeys. The German kings, even more than other medieval rulers, could only make their authority respected in the far-flung regions of their kingdom by traveling ceaselessly from duchy to duchy, from frontier to frontier. Wherever they stayed, their jurisdiction superseded the standing power of dukes, counts, and advocates, and they could collect the profits of local justice and wield some control over it. As they came into each region, they summoned its leaders to attend their solemn crown wearings, deliberated with them on the affairs of the Reich and the locality, presided over pleas, granted privileges, and made war against peacebreakers at home and on enemies abroad.

The promotion of the German church. The royal revenues came from the king's demesne lands and from his share of the tributes that Poles, Czechs, heathen Slavs, and Danes paid whenever he could enforce his claims of overlordship. There were also profits from tolls and mints that had not yet been granted away. The king's demesne was his working capital. He and his household lived on its produce during their wanderings through the Reich, and it also served to provide for his family, to found churches, and to reward faithful services done to him, especially in war. To swell the hosts, vassals had to be enfeoffed, and alienations were inevitable. The Salians, though they inherited the remains of Ottonian wealth as imperial demesne, brought little of their own to make up for its diminution. The last Saxon, Henry II (1002–24), and after him Conrad II, accordingly took to enfeoffing vassals with lands taken from the monasteries. Since the beneficiaries were often already powerful and wealthy men in their own right, no class of freeborn, mounted warriors linked permanently with the crown sprang from the loyalties and rewards of one or two reigns. In any case, the lion's share of grants went to the German church.

The growth of central government

From the Carolingians, the German kings inherited their one and only institution of central government: the royal chapel, with the chancery that does not seem to have been distinct from it. Service there became a recognized avenue of promotion to the episcopate for highborn clerks. In the 11th century, bishops and abbots conducted the affairs of the Reich much more than the lay lords, even in war. They were its habitual diplomats and ambassadors. Unlike Henry I, Otto I and his successors sought to free the prelates from all forms of subjection to the dukes. The king appointed most of them, and to him alone, as to one sent by God, they owed obedience.

Thus, there arose beside the loose association of stem lands in the German kingdom a more compact and uniform body with a far greater vested interest in the Reich: the German church. By ancient Germanic custom, moreover, the founder of a church did not lose his estate in the endowment that he had made; he remained its proprietor and protecting lord. The bishoprics, it is true, and certain ancient abbeys, such as Sankt Gallen, Reichenau, Fulda, and Hersfeld, did not belong to the king; they were members of the kingdom but under his guardianship. The greater churches therefore had to serve the rulers with mounted men, money, and free quarters. Gifts of royal demesne to found or to enrich bishoprics and convents were not really alienations but pious reinvestments, as long as the crown controlled the appointments of bishops and abbots. The church did not merely receive grants of land, often waste, to settle, develop, and make profitable; it was also given, as has been shown, powers of jurisdiction over its dependents. Nor did the kings stint the prelates in other regalian rights, such as mints, markets, and tolls. These grants broke up counties and to some extent even duchies, and that was their purpose: to disrupt the secular lords' jurisdictions that had escaped royal control.

This policy of fastening the church, a universal institution, into the Reich, with its well-defined frontiers, is usually associated with the name of Otto I, but it gathered momentum only in the reigns of his successors. The policy reached a climax under Henry II, the founder of the see of Bamberg in the upper Main valley; nonetheless, Conrad II, though less generous with his grants, and his son Henry III continued it. Bishops and abbots became the competitors of lay princes in the formation of territories, a rivalry that more than any other was the fuel and substance of the ceaseless feuds, the smoldering internal wars in all the regions of Germany for many centuries. The welter and the confused mosaic of the political map of Germany until 1803 is the not-so-remote outcome of these 10th- and 11th-century grants and of the incompatible ambitions that they aroused.

The Ottonian conquest of Italy and the imperial crown. Otto I's marriage with Adelaide (Adelheid), daughter of Rudolph II of Burgundy, and the Italian rivalries between his son Liudolf, duke of Swabia, and Otto's brother Henry I, duke of Bavaria, drew him southward. After 951, expeditions into Italy were a matter for the whole Reich under the leadership of its ruler and no longer just an outlet for the expansion of the south German tribes. For the Saxon military class, too, the south was more tempting than the primeval forests and swamps beyond the Elbe. With superior forces at their back, the German kings gained possession of the Lombard kingdom in Italy. There, too, their overlordship in the 10th and the 11th centuries came to rest on the bishoprics and a handful of great abbeys.

After his victory over the Magyars in 955, Otto I's hegemony in the west was indisputable. By the standards of one chronicler, the Saxon Widukind, he had already become emperor because he had subjected other peoples and enjoyed authority in more than one kingdom. But the right to confer the imperial crown, to raise a king to the higher rank of emperor, had fallen to the papacy, which had crowned Charlemagne and most of his successors. The Carolingian order in the west was still the model and something like a political ideal for all Western ruling families in the 10th century. Otto had measured himself against the political tasks that had faced his East Frankish predecessors and more or less mastered them. To be like Charlemagne, therefore, and to clothe his newly won position in a traditional and time-honoured dignity, he accepted the imperial crown and anointment from Pope John XII in Rome in 962. The substance of his empire was military power and success in war; but Christian and Roman ideas were woven round the Saxon's throne by the writers of his own and the next generation. Although the German kings as emperors did not give the law to the Roman church in matters of doctrine and ritual, they became its political masters for nearly a century. The imperial crown enhanced their standing even among the nobles and knights who followed them to Italy and can hardly have understood or wanted all its outlandish associations. Not only the king but also the German bishops and lay lords thus entered into a permanent connection with an empire won on the way to Rome and bestowed by the papacy.

The formation of the Holy Roman Empire

Otto II (reigned alone 973–983) and above all Otto III (983–1002) were strongly drawn toward their new Mediterranean sphere of action, but Henry II (1002–24) returned to a sober regime centred on Germany and contented himself with three brief Italian expeditions.

The Salians, the papacy, and the princes, 1024–1125. Under Conrad II (reigned 1024–39), the first member of the Rhine-Frankish house known as the Salians, the kingdom of Burgundy fell finally under the overlordship of the German crown, and this tough and formidable emperor also renewed German authority in Italy. His son and successor, Henry III (reigned 1039–56), treated the empire as a mission that imposed on him the tasks of reforming the papacy and of preaching peace to his lay vassals. Without possessing any very significant new resources of power, he gave to his authority an exalted and strained theocratic complexion. Yet, under him, the last German ruler to maintain his hegemony in western Europe, the popes themselves seemed to become mere imperial bishops. He deposed three of them, and four Germans held the Holy See at his command; but lay opposition to the emperor in Germany and criticism of his regime over the church were on the increase during the last years of his reign.

The papal reforms and the German church. More than any other feudal society in early medieval Europe, Germany was divided and torn by the revolutionary ideas and measures of the reformed papacy. Beginning with the pontificate of Leo IX (1048–54)—one of Henry III's nominees—the most determined and inspired spokesmen of ecclesiastical reform placed themselves at the service of the Holy See. Only a few years after Henry III's death (1056), they agitated against lay authority in the church, founded on proprietary rights. They regarded the laity as passive partakers of the sacraments and denied the supernatural status of kingship. Priests, including bishops and abbots, who accepted their dignities from lay lords and emperors at a price, according to the reformers, committed a sin; for these earthly powers could not rightly confer churches at all, nor could they own them. They believed, moreover, that thorough reforms could be brought about only by the exaltation of the papacy so that it commanded the obedience of all provincial metropolitans and was out of the emperor's and the local aristocracy's reach.

Effects of the reformed papacy

The endless repetition of the reformers' teachings in brilliant pamphlets and at clerical synods spread agitation in Italy, Burgundy, and Lotharingia—all parts of the empire. Their new program committed the leaders of the movement to a struggle for power because it struck at the very roots of the regime to which the German church had grown accustomed and on which the German kings relied. The vast wealth that Henry IV's predecessors had showered on the bishoprics and abbeys would, if the new teaching prevailed, escape his control and remain at the free disposal of prelates whom he no longer appointed. Under Roman authority the churches were to be freed from most of the burdens of royal protection without losing any of its benefits. The most fiery spirits in Rome did not flinch from the consequences of their convictions. Their leader Hildebrand, later Pope Gregory VII (reigned 1073–85), was ready to risk a collision with the empire.

Henry IV was not yet six years old when his father died in 1056. The full impact of the Gregorian demands—coming shortly after a royal minority, a Saxon rising, and a conspiracy of the south German princes—has often been regarded as the most disastrous moment in Germany's history during the Middle Ages. In fact, the German church proved thoroughly unreliable as an inner bastion of the empire even before Rome struck. Its leaders, Anno and Adalbert, archbishops of Cologne and of Hamburg–Bremen respectively, shamelessly exploited their hold over the young king by hunting for spoils out of the imperial demesne. In 1074 and 1075 Gregory proceeded against simony (the buying and selling of church office) in Germany and humiliated the aristocratic episcopate by summonses to Rome and sentences of suspension. These papal actions demoralized and shook the German hierarchy. The prelates' return to their customary support of the crown was not disinterested, nor wholehearted, nor unanimous.

Decrees of Gregory VII

The discontent of the lay princes. Henry IV's minority also gave elbowroom to the ambitions and hatreds of the lay magnates. The feeble regency of his mother, Agnes of Poitou, faltered before the throng of princes, who respected only authority and forces greater than their own. The ruling influence of the higher clergy at the court of Henry III and the renewed flow of grants to the church had estranged them from the empire. It is likely also that these eternally belligerent men were lagging behind the prelates in the development of their agrarian resources. The prelates had a vested interest in peace, and under royal protection they improved and enlarged their estates by turning forests into arable land and also by offering better terms to freemen in search of a lord. The bishops' market and toll privileges brought them revenues in money, which many of the lay princes lacked. So far, however, the princes' military power, their chief asset, had remained unchallenged. Now, for the first time, they also had to face rivals within their own sphere of action. Henry III and the young Henry IV began to rely on advisers and fighting men drawn from a lower tier of the social order—the poorer, freeborn nobility of Swabia and, above all, the class of unfree knights, known as *ministeriales.* These knights had first become important as administrators and soldiers on the estates of the church early in the 11th century. Their status and that of their fiefs was fixed by seignorial ordinances, and they could be relied on and ordered about, unlike the free vassals of bishops and abbots. Beginning with Conrad II, the Salian kings used *ministeriales* to administer their demesne, as household officers at court and as garrisons for their castles. They formed a small army, which the crown could mobilize without having to appeal to the lay princes, whose ill will and antipathy toward the government of the Reich grew apace with their exclusion from it.

Having come of age, Henry IV used petty south German nobles and his *ministeriales* to recover some of the crown lands and rights, which the lay princes and certain prelates had acquired during his minority, particularly in Saxony. His recovery operations went further, however, and a great belt of lands from the northern slopes of the Harz Mountains to the Thuringian Forest was secured and fortified under the supervision of his knights to form a royal territory, where the king and his court could reside. The south German magnates were thus kept at a distance when Henry and his advisers struck at such neighbouring Saxon princes as Otto of Northeim and the Billung family.

Formation of a royal territory

The storm broke in 1073. A group of Saxon nobles and prelates and the free peasantry of Eastphalia, who had to bear the brunt of statute labour in the building of the royal strongholds, revolted against the regime of Henry's Frankish and Swabian officials. To overcome this startling combination and to save his fortresses, the king needed the military strength of the south German princes Rudolf of Rheinfelden, duke of Swabia; Welf IV, duke (as Welf I) of Bavaria; and Berthold of Zähringen, duke of Carinthia. Suspicious and hostile at heart, they took the field for him only when the Eastphalian peasantry committed outrages that shocked aristocratic caste feeling everywhere. Their forces enabled Henry to defeat the Saxon tribal rebellion at Homburg near Langensalza in June 1075. But, when the life-and-death struggle with Rome opened only half a year later, the south German malcontents deserted Henry and, together with the Saxons and a handful of bishops, entered into an alliance with Gregory VII. Few of them at this time were converted to papal reform doctrines, but Gregory's daring measures against the king gave them a chance to come to terms with one another and to justify a general revolt.

The civil war against Henry IV. On Feb. 22, 1076, the pope absolved all men from their oaths to Henry and solemnly excommunicated him. In October Gregory's legates met the German lords at Tribur (modern Trebur) to decide on the future of the king, whom his last adherents now abandoned. Although Henry was absolved by Gregory at Canossa in January 1077, the princes two months later elected Rudolf of Rheinfelden to rule in his place.

The war that now broke out lasted for almost 20 years. A majority of the bishops, most of Rhenish Franconia (the Salian homeland), and some important Bavarian and Swabian vassals sided with Henry. He thus held a central position, dividing his south German from his Saxon enemies, who could not unite long enough to destroy him. With the death in battle of Rudolf of Rheinfelden (1080) and the demise of another antiking, Hermann of

Division of Henry's enemies

Salm (1088), the war in Germany degenerated into a number of local conflicts for the possession of bishoprics and abbeys. It almost died down in 1098, when the south German adherents of the papacy came to terms with Henry for the time being, though without recognizing his antipope Clement III. Throughout these years the crown, the churches, and the lay lords had to enfeoff more and more *ministeriales* in order to raise mounted warriors for their forces. Though this recruitment and frequent devastations strained the fortunes of many nobles, they knew how to recoup themselves by extorting more fiefs out of neighbouring bishoprics and abbeys. The divided German church thus bore the brunt of the costs of civil war, and it needed peace almost at any price.

Henry V and results of the conflict. The Salian dynasty and the rights for which it fought were saved because Henry IV's son and heir himself seized the leadership of a last rising against his father (1105). This maneuver enabled Henry V (reigned 1106–25) to continue the struggle for the crown's prerogative over the empire's churches against the inexorable demands of the papacy. The conflict now shrank into a legalistic dispute over the right to invest bishops and abbots with their dignities and the secular possessions attached to them. As the struggle continued, the princes became the arbiters and held the balance between their overlord and the pope. In 1122, acting as intermediaries and on behalf of the Reich, they forced the temporary concessions known as the Concordat of Worms out of the Holy See and its German spokesman, Archbishop Adalbert of Mainz, the bitter personal enemy of Henry V and the territorial rival of the Hohenstaufen sons of Henry's sister Agnes. By then, however, the princes had for the most part defeated efforts to restore royal rights in Saxony and to stem the swollen jurisdictions and territorial powers of the aristocracy elsewhere.

German loss of power after the death of Henry V

When Henry V, the last Salian, died childless in 1125, Germany was no longer the most effective political force in Europe. The brilliant conquest states of the Normans in England and Sicily and the patient, step-by-step labours of the French kings were achieving forms of government and concentrations of military and economic strength that the older and larger empire lacked. The papacy had dimmed the empire's prestige, and Rome became the true home of universalistic causes. When Pope Urban II preached the first crusade in 1095, Henry IV, cut off and surrounded by enemies, was living obscurely in a corner of northern Italy. The Holy See, by its great appeal to the militant lay nobility of western Europe, thus won the initiative over the empire. At this critical moment the Reich also lost control in the Italian bishoprics and towns, just when their population, trade, and industrial production were expanding fast. Germany did not even benefit indirectly from the crusaders' triumphs, although some of their leaders (*e.g.*, Godfrey of Bouillon and Robert II of Flanders) were vassals of the emperor. The civil wars renewed for a time the relative isolation of the central German regions.

Internally, the crown had saved something of the indispensable means of government in the control over the church; but it was a bare minimum, and its future was problematic. The ecclesiastical princes henceforth held only their temporal lands as imperial fiefs, for which they owed personal and material services. As feudatories of the empire, they came to represent the same interests toward it as did the lay princes; at least, their sense of a special obligation tended to weaken. The king's jurisdiction continued to exist alongside and in competition with that of the local powers. The great tribal duchies survived as areas of separate customary law. Each developed differently, and the crown could not impose its rights on all alike or change the existing social order. The most tenacious defenders of this legal autonomy had been the Saxons; but it also prevailed in Swabia, where distinct territorial lordships grew fast.

The Gregorian reform movement therefore aggravated the age-old contradictions in Germany's early medieval constitution, but its monastic culture and its intellectual interests were anything but barren. Both sides fought with new literary weapons to work on public opinion in cathedrals and cloisters and perhaps also in the castles of the lay aristocracy. In their hard-hitting polemical writings they attempted to expound the fundamental theological, historical, and legal truths of their cause. The agitation did something to disturb the cultural self-sufficiency of the German laity. It drove many of the south German nobles to maintain direct connections with the Holy See, and, whether they wanted to or not, they had to fall in with the aspirations of the religious leaders. The reform movement of the 11th and 12th centuries, it might almost be said, very nearly completed the conversion of Germany that had begun five centuries before.

GERMANY AND THE HOHENSTAUFEN, 1125–1250

Dynastic competition, 1125–52. The nearest kinsmen of Henry V were his Hohenstaufen nephews—Frederick, duke of Swabia (1105–47), and his younger brother Conrad, the sons of Henry's sister Agnes and Frederick, the first Hohenstaufen duke of Swabia. Some form of election had always been necessary to succeed to the crown, but, before the great civil war, nearness to the royal blood had been honoured whenever a dynasty failed in the direct line. By 1125, however, the princes, guided by Archbishop Adalbert of Mainz, no longer respected blood right. Affinity with Henry V was no recommendation to them, and hereditary succession seemed to lower their authority in the government of the Reich. Instead of Frederick they chose the duke of Saxony, Lothair of Supplingenburg (reigned as king 1125–37, reigned as emperor 1133–37). Like the Hohenstaufen, he had risen by a lucky marriage and a successful career of continuous fighting into the first rank of dynasts; but, unlike them, he had served the cause of the Saxon opposition to the Salians.

Disregard for blood right

With the enormous Northeim and Brunonian inheritances behind him, Lothair III (sometimes called Lothair II) could humble the Hohenstaufen brothers (1134) after marrying his only daughter and heiress to a Welf, Henry the Proud. Even without this dazzling alliance, the Welfs, already dukes of Bavaria and possessors of vast demesnes, countships, and ecclesiastical advocacies there, in Saxony, and in Swabia, were somewhat better off than their Hohenstaufen rivals. On the death of Lothair in 1137, however, the fears of the church and a few princes turned against the Welfs. Instead of Henry the Proud, who now held the duchies of Saxony and Bavaria and the Mathildine lands in Italy, they chose Conrad (reigned 1138–52), who had been Lothair's unsuccessful Hohenstaufen opponent.

The battle against the Welfs, which Conrad III put foremost on his political program, was abandoned with his death in 1152, when an election once again decided the succession and the political situation in Germany for the next 30 years. The princes then chose Frederick I Barbarossa (reigned as king 1152–90; as emperor 1155–90), the son of Conrad's elder brother Frederick and the Welf princess Judith. Frederick I agreed to share power in Germany with his Welf cousin Henry the Lion. The price of his election was dualism. In 1156 the duchy of Bavaria, which Conrad had tried to wrest from the Welfs, was restored to Henry the Lion, already undisputed duke of Saxony. The Babenberg margrave of Austria, Henry's rival, had to be compensated with a charter that raised his margravate into a duchy and gave him judicial suzerainty over an even wider area. Taken out of the Lion's duchy, it was to be held as an imperial fief that might descend both to sons and daughters. A perpetual principality, it served as a model for the aspirations of many other lay princes.

Colonization of the east. The history of Germany in the 12th and 13th centuries is one of ceaseless expansion. A conquering and colonizing movement burst across the river frontiers into the swamps and forests from Holstein to Silesia and overwhelmed the Slav tribes between the Elbe and the Oder. Every force in German society took part: the princes, the prelates, new religious orders, knights, townsmen, and peasant settlers. Agrarian conditions in the older lands of Germanic occupation seem to have favoured large-scale emigration. With a rising population, there was much experience in drainage and wood clearing but a diminishing fund of spare land to be attacked in the west. Excessive subdivision of holdings impoverished tenants and did not suit the interests of their lords. Some-

The conquering and colonizing movement of the 12th and 13th centuries

times also, seignorial oppression is said to have driven peasants to desert their masters' estates. They certainly found a better return for their labour in the colonial area: personal freedom, secure and hereditary leasehold tenures at moderate rents, and, in many places, quittance from services and the jurisdiction of the seignorial advocate.

The colonists brought with them a disciplined routine of husbandry, an efficient plow, and orderly methods in siting and laying out their villages. Very soon, even the Slav rulers of Bohemia and Silesia were competing for immigrants. First and foremost, however, the princes of the Saxon and Thuringian marches sought to attract settlers for the lands that they had conquered and the towns that they had founded to open up communications and trade routes. The older regions of the Reich, moreover, had not only peasants but also men of the knightly class to spare—soldiers who needed fiefs and lordships to uphold their rank. Both could be gained beyond the Elbe under the leadership of successful princes. The Germanized east thus became the home of fair-sized principalities in the 13th century, while all along the Rhine River valley the rights of government were tending to be scattered over smaller and less compact territories. The Ascanian dynasty, for instance, which under Albert the Bear began to advance into Brandenburg, by 1250 not only ruled over a broad belt of land up to the Oder River but had already established itself on the eastern banks ready for further advances. Farther south the Wettin margraves of Meissen busied themselves with settlements and town foundations in Lusatia.

For a time Henry the Lion, as duke of Saxony (1142–80), overshadowed all these rising powers, and the Welf profited as much by the ruthless use of his resources against weaker competitors as by his own efforts in Mecklenburg. As his was the only protection worth having in northeastern Germany, the newly established Baltic bishoprics were at his mercy, and he alone could attract the traders of Gotland to frequent the young port town of Lübeck, which he extorted from one of his vassals in 1158.

The Reich, too, possessed demesnes in the east, notably the Egerland, Vogtland, and Pleissnerland in the Thuringian March. The Hohenstaufen kings therefore took some part in opening up these regions. They, too, founded towns and monasteries on their thickly wooded lands and established their *ministeriales* as burgraves and advocates over them. But in this, as in many other things, they only competed with the princes. They did not and could not control the eastward movement as a whole.

Hohenstaufen policy in Italy. In the other great field of German expansion in the 12th century—Lombardy and central Italy—the emperors and their military following

Adapted from J.R. Strayer and D.C. Munro, *The Middle Ages, 395–1500*, 5th ed., p. 521, copyright © 1970; by permission of Appleton-Century-Crofts, Educational Division, Meredith Corporation



German expansion to the east, 800–1400.

alone counted. The rural population of Germany had no direct interest in the wars waged to recover and exploit regalian rights over the growing Lombard city communes. The connection between the German crown, the empire, and dominion over Italy has indeed been regarded as a disaster for Germany, and the ever-increasing concern of the Hohenstaufen dynasty with the south as its most tragic phase. Although Frederick Barbarossa's policy was opportunistic, he had really very little choice. Having bought off the Welfs and reconciled other great families with yet more concessions and lastly endowed his own cousin, Conrad III's son Frederick, with Hohenstaufen demesnes in Swabia, he had to try to mobilize their goodwill for the empire while it lasted. He now aimed to set up a regime of imperial officials and captains who were to exact dues and to control jurisdiction that the communes had usurped from the failing grasp of their bishops. The Germans in Italy did not bring valuable accomplishments to poor and primitive tribesmen, but they attacked economically advanced and better developed communities, to which they had nothing to offer in return for the rights and taxes they demanded. Military power was their chief asset in Lombardy, and they used it ruthlessly.

The Italian policy

For the Hohenstaufen *ministeriales* the rule of their masters in northern and central Italy was a career. Because they could be deployed continuously, they became the backbone of the imperial occupation. A handful of minor dynasts also served Barbarossa for many years in the powerful and profitable commands that he established. The German bishops and certain abbots still had to supply men and money, and some of them threw themselves wholeheartedly into the war: for instance, Rainald of Dassel and Philip of Heinsberg, archbishops of Cologne from 1159 to 1167 and from 1167 to 1191 respectively, who, as archchancellors for Italy, had a vested interest in it. The support of the lay princes, conversely, was fitful and sporadic. Even at critical moments they could not be counted on unless they individually agreed to serve or to send their much-needed contingents for a season. The refusal of the greatest of them, Henry the Lion, in 1176 brought about the emperor's defeat at the Battle of Legnano and spoiled many years' efforts in Lombardy.

The fall of Henry the Lion and the estate of princes. Forced to retreat before the papacy and the Lombard League in 1177, Barbarossa cooled toward his Welf cousin, whom he could justly blame for some of his setbacks. Dualism in Germany had outlived its purpose. Hitherto, the enemies of Henry—the princes, bishops, and magnates of Saxony—had been unable to gain a hearing against him at the emperor's court days. By 1178, however, the emperor was ready to help them. Outlawed (1180), beaten in the field and deserted by his vassals, Henry had to surrender and go into exile in 1182. His duchies and fiefs were forfeited to the Reich.

Surrender and exile of Henry the Lion

His fall left a throng of middling princes face to face with an emperor whose prestige, despite reverses, stood high and whose resources had greatly increased since he began to reign. The princes were nonetheless the chief and ultimate gainers from the events of 1180. The final judgment by which Henry the Lion lost his honours was not founded on folk law but on feudal custom. The princes who condemned him regarded themselves as the first feudatories of the empire, and they decided on the redistribution of his possessions among themselves. During the 12th century the stem duchies of the Ottonian period finally disintegrated. Within their ancient boundaries not only bishops but also lay lords succeeded in eluding the authority of the dukes. In their large immunities they themselves wielded stem-ducal powers. To enforce the imperial peace laws became both their ambition and their justification. Everywhere the greater lay dynasties and even some bishops tried to acquire a ducal or an equivalent title that would enable them to consolidate their scattered jurisdictions and, if possible, to force lesser freelords to attend their pleas.

These highest dynasts had interests in common, and they closed their ranks not only against threats from above but also against fellow nobles who had been less successful in amassing wealth, counties, and advocacies and who did not possess the superior jurisdiction of a duke, a margrave, a count palatine, or a landgrave. They and they alone were now called princes of the empire. To lend a certain cohesion to their varied rights, they were willing to surrender their houselands to the Reich and receive them back again as a princely fief. For the emperor it was theoretically an advantage that men so powerful in their own right should owe their chief dignity and most valued privileges to his grant. It opened the possibility of escheats (reversions), for in feudal custom the rules of inheritance were stricter than in folk right. In Germany, however, the political misfortunes of rulers succeeded, by and large, in ensuring that ancient caste feeling and notions of inalienable right conquered the principles of feudal law. By 1216 it was established that the emperor could neither abolish principalities nor create princes at random.

The "heirs" of Henry the Lion had to fight a ceaseless battle to establish and maintain themselves. In Bavaria the Wittelsbachs had received the vacant duchy, but they were not recognized as superiors by the dukes of Styria or by the dukes of Andechs-Meran. In Saxony the archbishop of Cologne was enfeoffed with Henry the Lion's ducal office and with all his rights in Westphalia, while an Ascanian prince, Bernard of Anhalt, received the eastern half of Henry's duchy. Neither Bernard nor the archbishop, however, could make much out of their dukedoms, except in the regions where they already had lands and local jurisdictions. All over the Reich these and regalian rights, such as mints, fairs, tolls, and the right of granting safe-conducts, were the substance of princely power. To possess them as widely as possible became the first goal of the abler bishops and lay lords.

The Hohenstaufen conflict with the papacy, 1159–1215. The attempt to establish a direct imperial regime in Italy antagonized the papacy once again and led to a new struggle with Rome, the ally of the Lombard communes. Political and territorial rather than ecclesiastical interests were at stake; but the popes could only fight as heads of the universal church, defending its liberty against a race of persecutors, and they had to employ their characteristic weapons—excommunication, propaganda, and intrigue. Nonetheless, the German bishops stood by Barbarossa and, for the most part, followed him in maintaining a prolonged schism against Pope Alexander III. Unsuccessful in Lombardy, the centre of Hohenstaufen ambitions after 1177 shifted to Tuscany, Spoleto, and the Romagna. This redoubled the fears and the resentment of the popes, particularly after 1189 when Frederick's son and chosen successor, Henry VI (reigned 1190–97), became the legitimate claimant to the Sicilian kingdom through his wife Constance, the sole surviving legitimate heiress.

With their backs to the wall, the popes had to make what use they could out of any opposition to the Hohenstaufen. Their chance came in 1197 when Henry VI died prematurely, leaving a three-year-old son, Frederick, to succeed him. To escape the chaos of a minority regime, the bulk of the German princes and bishops in 1198 elected the boy's uncle Philip of Swabia; but an opposition faction in the lower Rhenish region, led by the archbishop of Cologne and financed by Richard I of England, raised an antiking in Otto IV, a younger son of Henry the Lion. Pope Innocent III had to enlarge on his rights over imperial coronations and become a partisan in the German electoral feud if he wished to defend his recovered holdings in Italy against Hohenstaufen claims. Territorial interests in the Romagna tempted the papacy to exploit the weaknesses of the empire's constitution, the uncertainties of electoral custom, and the lack of strict legal norms in Germany. During the war for the crown, much hard-won demesne and useful rights over the church had to be sacrificed by the rivals to bribe their supporters.

Civil war

Frederick II and the princes. Henry's son Frederick II entered Germany to regain his own against Otto IV in 1212 and secured the crown in 1215. Despite promises to divide his inheritance, he kept the kingdom of Sicily and the empire together, and thus he also became locked in the inevitable life-and-death struggle with the papacy. The Hohenstaufen demesne in Swabia, Franconia, and Alsace and on the middle Rhine was still very considerable, and

Frederick even recovered certain fiefs and advocacies that had been lost during the earlier civil wars. Their administration was improved, and they provided valuable forces for his Italian wars. The great peace legislation of 1235, moreover, showed that the emperor had not become a mere competitor in the race for territorial gain. But, except for brief intervals, the princes and bishops were left free to fight for the future of their lands against one another and against the intractable lesser dynasts who refused to accept their domination. The charters that Frederick had to grant to the ecclesiastical princes (the so-called *Confoederatio cum principibus ecclesiasticis,* 1220) and later to all territorial lords (*Constitutio,* or *Statutum in favorem principum,* 1232) gave them written guarantees against the activities of royal demesne officials and limited the development of imperial towns at the expense of episcopal territories. But the charters were not always observed, and until 1250 the crown remained formidable in southern Germany, despite the antikings Henry Raspe and William of Holland, whom the papacy caused to be elected by the Rhenish archbishops in Germany in 1246 and 1247.

The Reich after the Hohenstaufen catastrophe. Frederick II died in 1250, in the midst of his struggle against Pope Innocent IV. His son Conrad IV left the north in 1251 to fight for his father's Italian possessions. William of Holland, antiking from 1247 to 1256, was thus without a rival in an indifferent Germany that had lost interest in its rulers. The bishops' cities and the towns, many of them founded on royal demesne, could not be absorbed. Their economic power challenged the age-old aristocratic order in German society, and, deprived of royal protection, they banded together to defend their autonomy. Within the nobility each rank tended to acquire some of the personal rights of its betters. The Hohenstaufen breakdown after 1250 left a gap in Swabia that no rising territorial power was able to fill. Countless petty lords and imperial *ministeriales* of the southwest succeeded in holding their seigniories as immediate vassals of the Reich. Their independent territories often survived for centuries.

Rise of the *ministeriales*

The *ministeriales* elsewhere, too, ceased to be the dependable servants that they once had been. Many free nobles voluntarily joined their ranks, and the knights thus assimilated the rights of the free aristocracy. They became the governing class of the territorial principalities, the standing councillors of their masters, whose household offices and local justice they monopolized and held in fee for many generations. Without the consent of this territorial nobility, the princes could neither tax nor legislate. Even the less important *ministeriales,* who only administered manors for their lords, entrenched themselves as hereditary bailiffs, who kept surplus produce for themselves and usurped seignorial dues, so that it paid the owners to commute the labour services of their villeins into money rents and to lease out those portions of the demesne that the unfree peasants had cultivated for them. Even then, however, the hereditary officials could not be easily dislodged. Finally, the ambitions of the princes themselves did not aim above the patrimonial policies of the past. They were acquisitive and attempted to build up their territories by usurpation, inheritance, marriage treaties, and escheats. They also tried, where possible, to administer their lands with officials whom they could depose at will. Yet they did so not to found sovereign states but chiefly to provide for their families. Again and again, they divided their dominions among sons, who, in turn, founded cadet lines and set them up on a fraction of the principality.

The loss of effective central authority

By 1250 there was thus no really effective central authority left in Germany. The prince-bishoprics had become fiercely contested prizes between neighbouring dynasties, often vassals of the see (*i.e.,* the bishopric). But constant feuds, disorder, and insecurity did not, by any means, frustrate the immense energies of the Germans in the 13th century. Eastward expansion continued under the leadership of the princes and, above all, of the Knights of the Teutonic Order. Their advance into Prussia went hand in hand with the opening up of the Baltic by the merchants of Lübeck. It is possible that three centuries of complete security from foreign invasion made it unnecessary for the German aristocracy to learn the virtues of political self-discipline and subordination; but it would be a great mistake to judge Hohenstaufen Germany solely by its failure to achieve political and administrative unity. (K.J.L.)

Germany from 1250 to 1493

1250 TO 1378

The extinction of the Hohenstaufen dynasty. The death of Frederick II in 1250 and of his son Conrad IV in 1254 heralded the irreversible decline of Hohenstaufen power in Germany and in the conjoint kingdoms of Naples and Sicily. Conrad's infant son Conradin, heir to Naples and Sicily, remained in Germany under the guardianship of his Bavarian mother. His uncle Manfred seized the reins of government in both Italian kingdoms and in 1258 formally supplanted Conradin by engineering his own coronation in Palermo. Manfred's defiance of papal claims to suzerainty over the kingdoms impelled the French-born Pope Urban IV to grant them to Charles of Anjou, brother of Louis IX of France. Papal taxation of the French clergy and loans from Florentine bankers enabled Charles to raise a large mercenary army for an expedition to Italy. Manfred, deserted by his barons, was defeated and slain near Benevento in 1266. Conradin then rallied his German supporters and led them across the Alps. But Conradin's financial resources were inadequate; unpaid troops deserted, and his depleted following was routed by Charles near Tagliacozzo (1268). Conradin was captured as he fled toward Rome, convicted of lese majesty (a form of treason), and beheaded in the public square at Naples.

Decline of Hohenstaufen power

The Great Interregnum. In Germany, the death of Frederick II ushered in the Great Interregnum (1250–73), a period of internal confusion and political disorder. The antikings Henry Raspe (landgrave of Thuringia, 1246–47) and Count William of Holland (ruled 1247–56) were elected by the leading ecclesiastical princes at the behest of the papacy. William's title was recognized initially only in the lower Rhineland, but his marriage to Elizabeth of Brunswick in 1252 ensured his acceptance by the interrelated princely dynasties of north Germany. The death of the Hohenstaufen Conrad IV left William without a rival in Germany. His growing strength and independence enabled him to escape from the tutelage of his ecclesiastical electors and to devote himself to purely dynastic policies. He pursued his feud with Margaret, countess of Flanders, over their conflicting territorial claims in Zeeland at the mouth of the Rhine. He renewed the attempts of his dynasty to obtain complete mastery of the Zuider Zee by thrusting eastward at the expense of Friesland; he died at the hands of the Frisians in 1256.

Pope Alexander IV forbade the election of a Hohenstaufen but interfered no further with the succession. Hence the initiative was taken by a small group of influential German princes, lay and ecclesiastical, acting out of self-interest. None desired the election of a ruler powerful enough to threaten their growing independence as territorial princes; nor did they single out a German candidate, who might prove to be as uncontrollable as William. Archbishop Conrad of Cologne approached Richard, Earl of Cornwall, brother of Henry III of England. Richard's gifts and assurances of future favour brought him the votes of the archbishops of Cologne and Mainz, the count palatine of the Rhine, and Otakar II of Bohemia. He was formally elected in 1257 and crowned king at Aachen (Aix-la-Chapelle). Three months after Richard's election, Alfonso X of Castile, who aspired to the empire in order to strengthen his foothold in Italy, was chosen in similar fashion by the archbishop of Trier, the duke of Saxony, the margrave of Brandenburg, and the devious Otakar.

Initiative of the German princes

The candidates were distracted by the turbulence of the aristocracy in their countries—Richard paid four fleeting visits to Germany; Alfonso failed to appear at all. Both appealed to the papacy for confirmation of their election. Their claims were summarized in Urban IV's bull *Qui Coelum* (1263), which assumed that the exclusive right of election lay with the seven leading princes involved in the double election of 1257.

The rise of the Habsburgs and Luxembourgs. *Rudolf of Habsburg.* When Richard died in 1272 the electoral

princes were spurred into action by Pope Gregory X, who desired the election of a German monarch sympathetic toward a crusade for the recovery of the Holy Land. The princes, dreading an overly powerful king, rejected the advances of Philip III of France and Otakar. They chose instead Rudolf of Habsburg (1273), a minor count of Swabia who lacked the strength to regain the crown domains the electors had usurped during the Interregnum. Papal diplomacy persuaded Alfonso X to abandon his pretensions to the throne; but Otakar denounced the election on the ground that the duke of Bavaria had voted as lay elector in his stead. Rudolf allied himself with the Wittelsbach family of Bavaria and with other envious neighbours of Otakar, who was defeated and slain (1278). The duchies of Austria and Styria, overrun by Otakar during the Interregnum, were declared vacant and conferred jointly on Rudolf's sons Albert and Rudolf (1282). These acquisitions placed the Habsburgs in the first rank of the German territorial princes and lent impetus to a gradual shift in the political centre of gravity from the Rhineland to east Germany. The growing Habsburg power, however, disquieted the electoral princes, who frustrated the king's attempts to secure the election of his elder son Albert in 1287 and of his younger son Rudolf in 1290.

Growing Habsburg power

Adolf of Nassau. On the death of Rudolf in 1291, the electors averted the danger of a hereditary Habsburg monarchy by choosing Count Adolf of Nassau as his successor. Adolf, possessing only a small patrimony to the south of the river Lahn, strengthened himself financially by promising military aid to and receiving subsidies from both sides in the current Anglo-French war. He took possession of Meissen when the cadet branch of the Wettin dynasty died out, and he used his foreign subsidies to purchase Thuringia in 1295. He was thus able to adopt a more independent attitude toward his electors. On June 23, 1298, five of the electors pronounced Adolf unfit to rule and deposed him; on the following day they elected Albert of Austria in his stead. Albert marched westward from Austria at the head of a large army, and in a battle at Göllheim, Adolf was slain and his supporters fled.

Albert I of Habsburg. By restoring the Habsburg Albert I (ruled 1298–1308) to the kingship, the electors placed themselves in jeopardy. The new ruler, backed by the ample resources of his Austrian dominions, was more powerful and unscrupulous than his predecessor. The electors regarded his treaty of friendship with Philip IV of France (1299) as a move to enlist French support for the election of his son Rudolf as his successor in Germany. His attempt to seize Holland and Zeeland as a vacant fief of the empire was rightly interpreted by the electors as an effort to establish Habsburg influence on the lower Rhine (1300). The four prince-electors of the Rhineland (the archbishops of Mainz, Trier, and Cologne and the count palatine) conspired to depose Albert. But Albert wrecked the design by decisive military action (1301–02), and he sealed his victory over the electors by obtaining confirmation in 1303 of his election from Pope Boniface VIII in return for an unprecedented oath of fealty and obedience to the papacy. Albert subsequently renewed Adolf's claims to Meissen and Thuringia, but his authority there was still disputed when he died by assassination in 1308. Albert temporarily tamed the electoral princes, placated the papacy, and renounced intervention in Italy; but this policy foundered at his death, and the electors were given a fresh opportunity to reassert their influence over the German monarchy.

Oath of fealty and obedience to the papacy

Henry VII of Luxembourg. The princes, released from Albert's heavy hand, sought a servant, not a master. Archbishop Baldwin of Trier sponsored the candidature of his brother, Count Henry of Luxembourg, who was elected at Frankfurt (modern Frankfurt am Main) in 1308 as Henry VII. The house of Luxembourg (Luxemburg) was not a major territorial power, and Henry lost no time in exploiting his new status to extend its possessions. Under his direction the Diet of Frankfurt (1310) closed the long-disputed question of the Bohemian succession by awarding the kingdom, with the consent of the Bohemian estates, to Henry's son John. Thus, in common with the Habsburgs, the main weight of Luxembourg interests gravitated eastward. But Henry, unlike his Habsburg predecessors, dreamed of a restoration of the ancient authority of the empire in Italy. His Italian expedition (1310–13) opened brilliantly, and in 1312 he was crowned Holy Roman emperor at Rome. The old fear of German domination, however, stiffened the resistance of the Italian states. Pope Clement V was alarmed by Henry's preparations to invade the kingdom of Naples, a papal fief, and threatened excommunication. A renewed collision of empire and papacy seemed imminent when Henry died in 1313.

The growth of territorialism under the princes. The decline of Hohenstaufen influence in Germany, the Interregnum, and the rapid alternation of dynasties on the German throne created favourable conditions for the territorial princes, lay and spiritual, to gain power. Frederick II had purchased the support of the princes by lavish grants of crown lands, chiefly in the Rhineland and Thuringia; in 1220 he procured the cooperation of the ecclesiastical princes in the election of his son Henry as king and eventual heir to the empire by renouncing his regalian rights of building castles, issuing coinage, and imposing tolls on merchandise in their territories. Henry himself had extended similar concessions to the lay princes in 1231.

Thenceforth the direct action of royal authority was virtually precluded in the princely domains. The princes were at liberty to multiply castles and toll stations, establish mints, exploit mineral deposits, and settle all judicial cases except those transferred on appeal to the court of the emperor. The machinery of administration under the prince and his council (*Hofrat*) was, nevertheless, still rudimentary. Public taxation was intermittent and restricted to emergency occasions, and it was subject to the consent of the three estates of the principality (clergy, nobles, townspeople), which were consulted separately by the prince. The estates grasped the opportunity to ventilate their grievances and to press their advice upon the prince. The emerging territorial state was thus under the dual government of the prince and the estates, and its development was to be heavily influenced by a shifting balance of power between them.

Downfall of royal authority in princely domains

Constitutional conflicts in the 14th century. The death of Henry VII led to a disputed election and a civil war in Germany. The electors' impulse to choose another lesser count as king was checked by the houses of Habsburg and Luxembourg, which pressured the prince-electors to choose between their candidates. The pro-Habsburg majority elected Frederick the Handsome, duke of Austria. The minority withdrew their support from Henry VII's son John and transferred it to a more formidable candidate, the Bavarian duke Louis of Wittelsbach, who had recently broken an Austrian invasion of his duchy.

Electoral custom did not yet acknowledge the majority principle. The papacy, which had claimed the right to adjudicate disputed elections since 1201, was vacant. Hence the two claimants settled their differences by the sword. In 1322 Louis defeated and captured his rival at Mühldorf, but his triumph in Germany merely raised the curtain on a long and bitter dispute with the papacy.

Pope John XXII, guided by canon law and precedent, affirmed that Louis might not legally rule until confirmed by the papacy; thus the disputed election of 1314 and the absence of papal approbation invalidated Louis' royal title and his right to govern. Louis contended, however, that election by a majority conferred a legitimate title and administrative power and did not require papal confirmation. His defiance of the pope exposed him to excommunication (1324) and to the procedures of canon law, whereby he was required to submit entirely to the papal terms before absolution could be granted. Louis warned the electors that their rights were endangered by the subjection of the elections to papal confirmation. Six electors responded in the Declaration of Rhens (1338), proclaiming as an ancient custom of the empire that election by a majority was valid and that the king-elect assumed his administrative power immediately, without the intervention of papal approbation. Under Louis' direction the declaration was repeated at the subsequent Diet of Frankfurt as an imperial law, and offenders against it were declared guilty of lese majesty.

The Declaration of Rhens

John XXII and his successors were unyielding. In 1343 Pope Clement VI made diplomatic overtures to Charles of Luxembourg, heir to the Bohemian throne, with the object of procuring his election to the German kingship in Louis' stead. The electors, led by Baldwin of Luxembourg, archbishop of Trier, began to desert Louis one by one. The pope thereupon urged a new election. Charles assured the pope secretly that he would await papal confirmation of his forthcoming election before exercising governmental power in the Italian possessions of the empire, but, despite intense pressure by Clement, he would accept no such restriction with regard to Germany. In 1346 only two electors remained faithful to Louis: his son Louis of Brandenburg and his kinsman Rudolf, count palatine of the Rhine. The other five assembled at Rhens on July 11 and elected Charles under the title of Charles IV. The new king was spared a lengthy conflict with his rival, who died of apoplexy in 1347.

Charles IV and the Golden Bull. Charles IV (ruled 1346–78) readily perceived that disputed elections exploding into civil war had been a standing malady of the German body politic since 1198 and that the stability of the German monarchy depended largely upon the degree of cooperation achieved with the territorial princes, more especially with the prince-electors. On his return from his imperial coronation as Holy Roman emperor (1355) he promulgated, with the consent of the German assembly of estates or diet (1356), a basic constitutional document, known as the Golden Bull from its pendant gold seal (*bulla*). Charles's double objective was to minimize areas of dispute in future elections and to strengthen his ties with the electors. Unanimity among the electoral princes had always been difficult to attain; hence the validity of election by majority vote, a principle already set forth in the Declaration of Rhens, was reaffirmed. The territories of the lay electors were declared indivisible and heritable only by the eldest son. Thus, partitions of land by family agreement and consequent uncertainty concerning the holder of the electoral vote were eliminated. In conformity with ancient custom the archbishop of Mainz was to convene the electors and to request them to name their favoured candidate. He was to announce his own choice after the other electors had given their vote verbally, and so he could cast the deciding vote in the event of a tie. The election was to be held in Frankfurt, the royal coronation in Aachen.

Provisions of the Golden Bull

The membership of the electoral body was fixed at the traditional number of seven: the archbishops of Mainz, Cologne, and Trier, the count palatine of the Rhine, the king of Bohemia, the margrave of Brandenburg, and the duke of Saxony. When the throne was vacant the count palatine would be regent in south Germany and the duke of Saxony in the north; thus the long-standing papal claim to govern the empire during a vacancy was tacitly rejected. The question of papal confirmation of elections was ignored; neither Charles nor his electors were prepared to yield, but an open affirmation of their position would have been ill-received by the papacy, which had played a leading role in Charles's election.

The Golden Bull consolidated and extended the territorial power of the electors. Their right to construct castles, issue coinage, and impose tolls was confirmed. They could judge without appeal. Conspiracy or rebellion against them was deemed high treason. They were to meet the ruler once yearly as supreme advisory council on affairs of state. The formation of city leagues against them was specifically prohibited. On the basis of these enactments the Golden Bull has been called the Magna Carta of German particularism. The electors in their capacity of territorial lords were its chief beneficiaries; the rest of the princes were envious and strove thenceforth to acquire an equally large measure of territorial sovereignty.

Rudolf IV of Austria ordered his chancery to fabricate a series of imperial charters, including two from Julius Caesar and Nero, as evidence of his virtual independence of the empire. Charles IV submitted them for examination to the Italian humanist Petrarch, who declared the charters to be spurious. Rudolf took up arms and was bought off by the recognition of his claim to Tirol (1364).

The election of Charles's son Wenceslas (Wenzel) as king in 1376 (two years before Charles's death) was a striking example of the emperor's skill in securing the cooperation of the electors for his dynastic purposes. The election of an emperor's son as king of the Romans during the father's lifetime had not occurred since 1237; the prince-electors, in their anxiety to prevent any single dynasty from strengthening its grip on the succession, had checked all subsequent attempts. But unprecedented gifts, concessions, and a renewed prohibition of city leagues by Charles overcame the opposition of the electors. Pope Gregory XI had previously announced that the election would be invalid without papal confirmation. Charles, in concert with the electors, speeded the election and subsequent coronation of his son and then submitted an antedated request for confirmation to the pope, who countered these devious tactics by delaying confirmation; it was still under consideration at Gregory's death in 1378. The decline of the papacy during the Great Schism (1378–1417) precluded the vigorous assertion of its right of confirmation, which became a mere formality and was subsequently tacitly abandoned.

Election of Wenceslas

Decline of the German monarchy. Charles IV's power was based primarily upon the territorial possessions of the house of Luxembourg, which he greatly extended by the purchase of the electorate of Brandenburg (1373). The German monarchy was a source of dignity and influence, but in terms of land and revenue it was outranked by Charles's hereditary domains in the east and northeast. The Golden Bull, replete with privileges to the electors, attacked none of the fundamental problems of the monarchy: dwindling crown lands, slender revenues, lack of an army and of an expert bureaucracy.

The financial problem was acute and of long standing. The succession of disputed elections between 1198 and 1257 had compelled the various claimants to purchase support by grants of royal land and revenues; the attempt by Rudolf of Habsburg to recover possession of crown lands alienated since 1245 had been opposed by his electors, who were unwilling to set an example by surrendering their own considerable acquisitions. At every election the votes of the princes had been secured by the grant or pledge of royal rights and property; thus, every king began his reign with a financial millstone round his neck and could attain freedom of action only by the possession or acquisition of extensive dynastic territories. The system of pledging crown lands involved the transference of the land and its revenues to the creditor. These revenues did not reduce the original debt, and the alienation tended to become permanent. The imperial cities (*Reichsstädte*) had been heavily taxed by Rudolf, and before his acquisition of Austria they had furnished the bulk of his revenue. His less provident successors had pledged them in a few cases to the local territorial princes and had thus lost the right of taxation. Charles IV carefully cultivated his dynastic revenues from Bohemia, but he lavishly expended crown assets in Germany to expand his family possessions. His financial exploitation of the cities for purely dynastic purposes naturally stiffened their resistance to taxation. By 1400 the annual revenues from all the German crown possessions averaged only 30,000 florins.

Financial problems

The enforcement of the public peace, a taproot of royal power in other countries, had long since slipped from the hands of the German monarchs. The German monarchy possessed no executive officials comparable with the English sheriff or justice of the peace, and it was diverted from its guardianship of law and order by recurrent conflicts with the papacy and by its absorption in purely dynastic matters. Consequently, the proclamation and enforcement of the peace fell into the hands of regional associations of cities and of the individual territorial princes. Thus the monarchy was prevented from using its function as defender of the public peace as an entering wedge to invade the jurisdiction of the municipalities and the territorial lords.

In sum, the German rulers were being gradually deprived of their triple role of feudal suzerains, defenders of the church, and keepers of the peace. The sweeping privileges granted to the princes in 1220 and 1231 had undermined

their position as feudal suzerain. Their bitter struggles with the papacy cast doubt on their credibility as protectors of the church. They allowed their powers as guardians of the public peace to slip into the hands of others.

The continued ascendancy of the princes. By Charles IV's death in 1378, the division of Germany into loosely defined territorial principalities had reached an advanced stage.

Division into territorial principalities

Southern Germany. In south Germany the dissolution of the Hohenstaufen duchy of Swabia gave territorial predominance to the Habsburgs, whose original possessions lay in Alsace, Breisgau, the Vorarlberg, and Tirol. Rudolf's acquisition of Austria and Styria (1282) had more than doubled the Habsburg patrimony and established its centre of gravity in eastern Germany. The Habsburg's rivals and neighbours to the north, the counts of Württemberg, had combined with the Swabian nobles to foil the attempt of Rudolf to revive the defunct duchy of Swabia for one of his sons (the counts, insatiably acquisitive and the inveterate enemies of the cities of the region, were finally raised to ducal status in 1495). The margraves of Baden were chiefly preoccupied with the southward expansion of their territory on the upper Rhine at the expense of the independent small nobles and cities of Swabia.

These three large entities contained lesser lordships, which were in constant danger of absorption by marriage, purchase, or feud. Bavaria, granted to the house of Wittelsbach as a duchy in 1180, was strengthened by the acquisition of the Palatinate in 1214; but subsequent testamentary partition restricted this important gain to the Upper Palatinate.

Central Germany. In central Germany the margraves of Meissen of the Wettin dynasty thrust steadily eastward and received the electorate of Saxony in 1423, when the Ascanian line of electors died out; in the west they obtained Thuringia (1263) and clung to it tenaciously despite repeated royal attempts to oust them by claiming it as a vacant fief. The landgraves of Hesse, though surrounded by powerful neighbours, contrived to make modest territorial gains at the expense of the Wettin dynasty and the archbishops of Mainz. East and south of Hesse, the Rhine–Main region was a land of great ecclesiastical princes: the archbishops of Mainz, Trier, and Cologne; the bishops of Speyer, Worms, Würzburg, and Bamberg; and the wealthy abbots of Fulda and Lorsch. It abounded in counts of the second rank, dominated by a great secular prince, the count palatine of the Rhine. The area contained four electorates and was therefore of crucial political importance.

Northern Germany. In north Germany the dukes of Brunswick dissipated their strength by frequent divisions of their territory among heirs. Farther east the powerful duchy of Saxony was also split by partition between the Wittenberg and Lauenburg branches; the Wittenberg line was formally granted an electoral vote by the Golden Bull of 1356. The strength of the duchy lay in the military and commercial qualities of its predominantly free population. But the vigour of its eastward expansion into the Slav lands beyond the Elbe tended to diminish its involvement in the internal politics of the Reich.

Eastern Germany. In eastern Germany the duchy of Mecklenburg, Germanized by a steady stream of immigrants, was drawn deeply into Scandinavian affairs and in 1363 provided Sweden with a new royal dynasty in the person of Albert of Mecklenburg. The electorate of Brandenburg, purchased by Charles IV and bequeathed to his second son, Sigismund, was dominated by a disorderly and rapacious nobility. Sigismund granted this dubious asset in 1415 to his faithful ally Frederick, burgrave of Nürnberg. The kingdom of Bohemia remained the durable territorial core of the Luxembourg dominions, and its silver mines at Kuttenberg, under German supervision, vastly increased crown revenues. The Slav population resented increasingly the economic and cultural influence of the German minority, and this created antagonisms profoundly disturbing to the monarchy.

Continued dispersement of territory. Inside the various territories the consolidation of the princely authority was far from complete. The principalities were often ragged in outline and territorially dispersed because of the accidents of inheritance, grant, partition, and conquest. Everywhere lesser nobles disputed the power of the prince and formed associations in defense of their rights and fiefs. In the ecclesiastical princedoms the ascendancy of an archbishop or a bishop was contested by the cathedral chapter, which had become a preserve of the nobility. The self-governing cities fought to protect their chartered liberties and drew together in formidable leagues to resist princely encroachment. Thus the princes, trying to enforce their authority, tended to consolidate the opposition and to excite potential or open hostility.

Opposition to princely power

In this crucial struggle the great secular potentates impaired their own strength by persisting in the Germanic custom of dividing their territory among their sons instead of transmitting it intact to the eldest. By 1378 the Bavarian lands of the house of Wittelsbach were shared between three grandsons of Louis IV. In 1379 the wide possessions of the Habsburgs were partitioned by family agreement between Albert III and his younger brother Leopold.

The ecclesiastical princes, vowed to celibacy and elected by their cathedral chapters, could not hand on their lands to their descendants. Still, their policies and aspirations were not much different from those of the secular princes, and most of them managed to install their relatives in rich canonries and prebends.

1378 TO 1493

Internal strife among cities and princes. The electors had voted for Wenceslas reluctantly during Charles IV's reign, fearful that the monarchy might become a perquisite of the house of Luxembourg. Most of the other princes shared their concern over the continued ascendancy of the dynasty.

Wenceslas. Wenceslas (ruled 1378–1400) inherited a variety of problems, which grew after his father's statesmanlike hand had been removed. Wenceslas' habitual indolence and drunkenness, vices that increased as he grew older, excited the indignation of his critics. His prolonged periods of residence in Bohemia betrayed his lack of interest in German affairs and allowed the continuous friction between princes, cities, and nobility to develop into open warfare.

Conflict between princes and cities

The collision of princes and cities was prompted by vital issues of long standing. The flight of the rural population from servile tenures on the land to the free air of the cities reduced the labour force and impaired the revenues of territorial lords. Others who stayed on the land accepted the protection and jurisdiction of the neighbouring city as "external" citizens (*Ausbürger, Pfahlbürger*) and thus withdrew themselves and their land from seignorial control. Only the most powerful cities (*e.g.,* Nürnberg, Rothenburg) were able to extend their extramural territory to a substantial degree by force, but all strove to expand the area of their jurisdiction at the expense of local lords, partly to prevent village industries from competing with the city guilds.

A second major issue was the insistence of territorial lords on imposing tolls on city merchandise in transit through their possessions. In theory, tolls on road and river traffic were exacted in return for the protection of merchants and their goods, but the multiplication of toll stations hampered trade and provoked innumerable disputes, which often culminated in the seizure of merchants and merchandise by exigent lords.

The third and immediate cause of the crisis lay in the financial policy of Wenceslas himself. His Bohemian revenues, though large, were strained by the great sums payable to the electors in return for his elevation to the kingship. Hence he attempted to tap the resources of the imperial cities by demanding heavy taxes, and he threatened to mortgage recalcitrant cities to the neighbouring princes, their capital enemies.

On July 4, 1376, an alliance of 14 imperial cities of Swabia was formed under the leadership of Ulm and Constance for mutual protection against unjust taxes and alienation from the empire. The Swabian League counted 40 members by 1385 and was linked with similar coalitions in Alsace, the Rhineland, and Saxony. Wenceslas' initial hostility to the league faded as its membership in-

creased, and in 1387 he gave it his verbal and unofficial recognition. He feared to offend the territorial princes by extending full recognition; further, a clause of the Golden Bull had declared all city leagues to be illegal. Thus he temporized and awaited the outcome of the approaching trial of strength between cities and princes. On Aug. 28, 1388, the princes of Swabia and Franconia routed the largely mercenary forces of the Swabian League at Döffingen, near Stuttgart. The stipendiaries of the Rhenish League were put to flight by the count palatine Rupert II near Worms on November 6.

The cities triumphantly withstood the ensuing siege operations, but their economy was injured by the forays, ambuscades, and blockades instituted by the princes. The protracted campaigns also exhausted the financial resources of the princes. When Wenceslas intervened in 1389, both parties were ready for peace. At the Diet of Eger (May 2) he ordered them to desist and declared the city leagues to be dissolved. The contestants complied. The princes were satisfied with the prospective disbandment of the cities, and the cities feared the consequences of further resistance, but Wenceslas' opportunism was relished by neither side. The princes disliked his political flirtation with the cities, and the cities resented his final championship of the cause of the princes.

Wenceslas' alienation of cities and princes

Wenceslas' early gestures of support for the cities rankled with the electors, who in 1384 and 1387 discussed the advisability of replacing him by an imperial vicar or regent. Wenceslas, however, learned of the plan and conveyed his opposition; nor could the electors unite on their choice of a regent. Some electors turned to a more drastic solution—his deposition. In 1394 Rupert II and the archbishop Frederick of Cologne considered the election of Richard II of England but failed to win the support of their electoral colleagues. In the following year, however, Wenceslas' elevation of Gian Galeazzo Visconti, imperial vicar of Milan, to the status of duke was assailed as a dismemberment of the empire and enabled the electors to act as the indignant defenders of the integrity of the Reich against a wasteful and profligate king. Wenceslas attempted to conciliate the princes by appointing his younger brother Sigismund as German regent (1396). But the Milanese issue enabled Rupert and Frederick to enlist the support of the archbishops of Mainz and Trier for their proposed deposition of Wenceslas. The death of Rupert in 1398 occasioned some delay. But at length the electors compiled a lengthy series of charges against the king, and in September 1399 they openly proclaimed their intention of deposing him.

At this critical stage further proceedings were temporarily checked by serious differences concerning the choice of Wenceslas' successor. The favoured candidate of the Rhenish electors was the count palatine Rupert III, himself an elector. But another elector, Duke Rudolf of Saxony, and a powerful group of north German princes contended that the electors could not raise one of their own members to the kingship. The Golden Bull had declared otherwise, but Rudolf held his ground and declined to participate in the subsequent proceedings. On June 4, 1400, the four Rhenish electors invited Wenceslas to Oberlahnstein to consider measures for the reform of the empire and threatened to release themselves from their oath of allegiance if he failed to appear. The king's efforts to rally support for his cause were utterly fruitless, and he decided to stay in Bohemia. On August 20 Archbishop John of Mainz, on behalf of the four electors, publicly proclaimed the deposition of Wenceslas as an unfit and useless king and freed his German subjects from their allegiance to him. On the following day the three archbishops elected Rupert in Wenceslas' stead. Rupert's consent to his election was presumed to furnish the necessary majority required by the Golden Bull.

Deposition of Wenceslas

Rupert. Rupert (ruled 1400–10) lacked the skill and resources necessary to revive the drooping power of the German monarchy. His title was not beyond dispute while Wenceslas lived, and the territorial princes and cities were therefore slow to acknowledge him. Pope Boniface IX, maintaining that only a pope might legally depose a German monarch, withheld his approbation of Rupert. An expedition against Wenceslas (1401) failed before the walls of Prague. Rupert then embarked upon an Italian expedition (1401–02), hoping to obtain the imperial crown from the pope and thus dispel the cloud of uncertainty that hung over his title. The enterprise was crippled by lack of financial means, Boniface's conditions were exorbitant, and Rupert returned to Germany without the coveted imperial coronation. Fortunately, he had little to fear from Wenceslas, who was fully occupied in protecting his Bohemian throne from the machinations of his ambitious younger brother Sigismund. Far more dangerous was the degeneration of Rupert's relations with the Rhenish electors. In 1405 he offended Archbishop John of Mainz by refusing him military aid in his war against Hesse and Brunswick. Consequently the archbishop united all the enemies of Hesse and Brunswick in the League of Marbach, which included 18 imperial cities. Rupert contended that coalitions of cities were prohibited by the Golden Bull, and he denounced the league as illegal. The dispute was arrested by the mediation of the archbishop of Cologne, but the memory rankled. Rupert's prospects darkened still further in 1408, when he lent his support to Pope Gregory XII against the cardinals who wished to summon a general council to end the Great Schism in the church. The archbishops of Mainz and Cologne and the vast majority of the German prelates favoured the conciliar solution and strongly approved the policy of the cardinals. Wenceslas shrewdly followed suit and in return received assurances from the cardinals that the future general council would recognize him as German king. The powerful proconciliar party in the German church proceeded to agitate openly for the restoration of Wenceslas to the throne. The threat of civil war, however, was averted by Rupert's death on May 18, 1410.

Sigismund. On the death of Rupert the movement for the reinstatement of Wenceslas immediately lost headway. The Rhenish electors, having deposed Wenceslas 10 years previously on ground of his unfitness, could not reelect him without admitting their inconsistency. Nonetheless, the House of Luxembourg was powerful and would assuredly throw its full weight against any non-Luxembourg candidate to the German throne. The four electors agreed on the expediency of selecting Rupert's successor from the Luxembourg dynasty but disagreed on the choice of candidate. Rupert's successor, the count palatine, and the archbishop of Trier elected Wenceslas' brilliant but unreliable brother, Sigismund, at Frankfurt on Sept. 20, 1410. Eleven days later, the archbishops of Cologne and Mainz elected Wenceslas' turbulent and treacherous cousin, Jost of Moravia. Jost died the next year, and Wenceslas agreed to accept Sigismund on condition that he himself retained the title of German king. But Sigismund ignored the reservation and assumed the disputed title. Wenceslas' protests were greeted with indifference in Germany and quickly died away. A second election of Sigismund at Frankfurt (July 21, 1411) gave him an ample majority and removed all doubt concerning the validity of the previous election.

Sigismund was energetic, versatile, and intelligent; but long experience never blunted his rashness in rushing into new projects, and his financial incapacity never ceased to astonish his contemporaries. His pursuit of personal power and dynastic possessions was unceasing and was conducted with complete unscrupulousness. His kingdom of Hungary and his later acquisition, Bohemia, were his primary concerns, and the interests of Germany were constantly set aside in their favour. The disastrous reigns of his predecessors, Wenceslas and Rupert, had emphasized Germany's basic problems: the weakness of the monarchy, the friction between princes and cities, and the unchecked growth of lawlessness and disorder. During his long reign (1410–37) Sigismund appeared less and less frequently in Germany and did little to correct these evils.

Sigismund's character

The Hussite controversy. Sigismund's prolonged absences were caused in great part by the explosive Hussite controversy in Bohemia. The Czech church in Bohemia had long retained a marked individuality and much autonomy in its liturgy. This independent temper in ecclesiastical affairs was being slowly fused in the late 14th century with a rising sentiment of nationality among the Czechs. The upsurge of feeling took the negative form of a grow-

ing hatred of the German minority, which dominated the towns by virtue of its economic power and cultural influence. The luxury and immorality of the Bohemian clergy were castigated by a series of religious reformers such as Conrad of Waldhauser, Thomas of Štítný, John Milíč of Kroměříž (Kremsier), and Matthew of Janov. The teachings of Conrad and Milíč assumed a strongly puritanical tinge; in opposition to the wealthy sacramental church with its external means of grace they held up the ideal of the primitive church in a condition of apostolic poverty and the exclusive authority of the Bible as the foundation stone of faith and belief. These three movements met and intermingled in the person of Jan Hus.

Jan Hus. A graduate in divinity of Charles IV's foundation, the University of Prague, Hus was appointed incumbent of the Bethlehem Chapel in Prague (1402) and immediately attracted wide attention by his sermons, which were delivered in Czech in accordance with the foundation charter of the chapel. In 1403 he strongly defended a number of extracts drawn from the religious writings of the Englishman John Wycliffe. Czech opinion in the university solidly supported Hus, but the more numerous German masters carried the day, and the teaching of the controversial extracts was forbidden. Similarly, when Pope Gregory XII's cardinals rebelled against the pope and demanded a general council to terminate the schism in the papacy (1408), the Czech members of the university aligned themselves with the cardinals, while the Germans stood with the pope. On matters of general policy the masters of the university voted by "nations," of which there were four: Bohemian, Bavarian, Saxon, and Polish, the last consisting in fact largely of Germans. Thus the Germans controlled three votes, the Czechs only one. When King Wenceslas reversed the proportion by decree, the German masters and students seceded to found their own university at Leipzig, and the mutual enmity deepened.

In 1410 Hus was excommunicated by Archbishop Zbyněk of Prague but refused to appear at the papal court for judgment and continued to preach. Two years later he protested against the sale of indulgences and was placed under papal sentence of excommunication. The city of Prague was subjected to a papal interdict. Hus left the capital at Wenceslas' request but preached throughout the land and vastly enlarged his following. From the lower Czech clergy who popularized Hus's doctrines, the masses learned that the German minority were intruders, foes of Bohemia and of the true religion. Lesser nobles who had lost their lands by mortgage or purchase to the prosperous German burghers of the cities were readily converted. The more self-interested members of the upper nobility were attracted by Hus's proposed reduction of the Czech church to apostolic poverty, which would bring the rich territorial possessions of the higher clergy within their grasp.

The Council of Constance

The rising ferment in Bohemia disquieted the heir apparent, Sigismund, and he intervened with the suggestion that Hus should expound and justify his opinions to the Council of Constance (1414–18), recently convened to heal the schism in the papacy. Hus accepted, and Sigismund furnished him with a comprehensive safe conduct. The conciliar commission that examined Hus focused the debate on two issues: the unauthorized Bohemian practice of extending communion in both kinds (bread and wine) to the laity, and the points of agreement between Wycliffe and Hus. Hus declined to retract his Wycliffite opinions until they were refuted by Holy Writ, and thus he defied the authority of the council in matters of doctrine. He was arrested on Nov. 28, 1414, and died at the stake on July 6, 1415. Sigismund's protests against the breach of his safe conduct were silenced by the argument that excommunicates automatically lost imperial protection.

The Hussite wars. The death of Hus enshrined him at once as a martyr and a national hero in the memory of his followers among the Czechs. They raised a storm of denunciation against Sigismund and expressed their resentment by widespread attacks on orthodox priests and churches. The Catholics retaliated in kind, and Bohemia was in a state of civil war when the death of Wenceslas (Aug. 16, 1419) brought Sigismund to the tottering throne. The new king's talent for conciliation and compromise was useless in the heated religious atmosphere. Pope Martin V urged him on against the Hussites and promised him imperial coronation as his reward. Under his prompting, Sigismund raised a motley host in Germany and launched it into Bohemia under the banner of a papal crusade (March 1, 1420). But the invaders were thrown back from the walls of Prague, and on July 7, 1421, Sigismund was declared deposed by the Bohemian assembly of estates. The shock of defeat forced Sigismund to attempt a fuller mobilization of German resources. Under the traditional system, princes and cities had been allowed to fix at their own discretion the quota of men provided by each when a royal campaign was in prospect. Naturally, both estates used their discretionary power to reduce contributions to a minimum. In 1422, however, Sigismund himself fixed the strength of the contingents demanded from the individual princes and cities throughout Germany. The response was disappointing. In 1426 the king raised his requirements, but to no effect. Hence the yearly campaigns against the Hussites were waged largely by mercenary armies. To meet the rising costs, the Diet of Frankfurt was persuaded in 1427 to vote a general tax, the so-called Common Penny. But there was little enthusiasm in Germany for the crusade; massive evasions of payment occurred, and the strength of local feeling hampered the coercion of defaulters.

Hussite raids into Germany

In 1429–30 the irrepressible Hussites swept through Saxony, Thuringia, and Franconia in a destructive foray. Sigismund, exploiting the general alarm, reverted to the older system and demanded contingents from each prince and city. The response improved, and a large army invaded Bohemia, only to meet complete disaster at Taus (Domažlice in modern Czechoslovakia) in 1431. It was evident that the veteran Hussites could not be crushed by force. Sigismund therefore welcomed the opportunity to transfer the problem of reconciling the Hussites with the church to the Council of Basel (1431–49). The Hussite extremists, the Taborites, were inflexible. They condemned the hierarchical system of church government and affirmed the priesthood of all true believers. Hence the council conducted its long and arduous negotiations with the majority party among the Hussites, the Calixtines or Utraquists, who were prepared to accept the grant of communion in both kinds as a basis of settlement. The Utraquist nobles annihilated the protesting Taborites at the Battle of Lipany (May 30, 1434), made peace with the council by the Compact of Iglau (July 5, 1436), which conceded them communion in both kinds, and reunited with the Roman Catholic church. The Utraquist nobles extracted far better terms from Sigismund as the price of their recognition. He agreed to accept the guidance of Czech councillors in governmental affairs, to admit only Czechs to public office, to grant an amnesty for all offenses committed since the death of Wenceslas, and to allow the Czechs a large measure of autonomy in their civil and religious life. It is unlikely that the slippery Sigismund intended to honour these pledges, but they cleared the way for his triumphant return to Prague in August 1436.

Effects of the wars on Germany

In Germany, the Hussite threat had clearly revealed the inadequacies of the existing financial and military systems, but the incentive to press Sigismund's reforms to a successful conclusion faded when the Hussite peril was scotched by negotiation. The general apathy was demonstrated in 1434, when Sigismund proposed to the princes a land peace embracing the whole of Germany. The abolition of private wars and feuds by such a peace was undeniably a paramount necessity. The princes themselves, however, were among the chief offenders against law and order, and their nominal approval of the plan deceived no one. Sigismund himself, increasingly absorbed in crucial negotiations with the Hussites, did not persevere, and the project gathered dust in the imperial archives. The impulse he gave to the cause of reform did not, however, fade entirely, though Sigismund did not live to see the sequel. His death on Dec. 9, 1437, terminated the tenure of the German throne by the House of Luxembourg and opened the door of opportunity to the Habsburg dynasty.

The Habsburgs and the imperial office. *Albert II.* In the

absence of a male heir, Sigismund had named his son-in-law Albert of Habsburg, duke of Austria, as his successor. Albert was able and vigorous, and the union of the territories of the two dynasties enabled him to exert considerable leverage in German politics. Albert declared his neutrality in the current dispute between Pope Eugenius IV and the Council of Basel on the subject of conciliar sovereignty and thereby evaded an issue on which the electors were strongly divided; thus, on March 18, 1438, he was unanimously elected at Frankfurt. The electors attempted to elicit from the new king an understanding that he would grant privileges to his subjects only with their advice and consent. They also submitted a project for the division of Germany into four new administrative units (*Kreise*) in which the enforcement of the land peace would be entrusted to captains of princely rank. Albert judged that the princes were seeking to enlarge their power and influence under the guise of introducing reforms for the common good. The German cities also doubted the impartiality of the princes as custodians of law and order. Both proposals were therefore stillborn. The king hastened from Frankfurt to defend his kingdom of Hungary, endangered by Turkish raids on Siebenbürgen (Transylvania in modern Romania). The campaign was brought to a premature close by the death of the king on Oct. 27, 1439.

Frederick III. Albert II had left only an infant son, and the leadership of the House of Habsburg passed to his cousin Frederick, duke of Styria. Inside the electoral college the duke was vigorously supported by his brother-in-law Frederick of Saxony and was elected unanimously on Feb. 2, 1440. The choice of Frederick tightened the hold of the Habsburgs on the German kingship. It also brought to the throne a ruler who, absorbed in dynastic concerns and in astrology, had no more than a passing interest in Germany. Under the absentee government of Frederick III, the feuds among the princes and the collisions between the princes and the cities developed into savage wars accompanied by widespread ravaging and pillage. All paid lip service to the need for peace; but who was to enforce it? Was it to be enforced by the monarchy, which lacked power and executive machinery? Was it to be enforced in the courts of the princes, whose judicial impartiality was suspect? Were complaints against the princes to be heard and decided in the king's court (*Hofgericht*)? Or must they be adjudicated by the council (*Hofrat*) of the prince concerned? The right to enforce peace effectively was a major source of power to the holder; hence the struggle between Frederick and the princes was long, bitter, and inconclusive.

The Turkish threat

These issues were brought to a head by the rapid westward progress of the Ottoman Turks after their victories at Varna on the Black Sea (1444) and at Kossovo in Serbia (1448). The Habsburg kingdom of Hungary and Frederick III's own duchy of Styria lay full in the path of the invaders. In 1453 the fall of Constantinople extinguished the Eastern Empire and aroused fears in Germany that the Western Empire would meet the same fate. The king used the opportunity to demand financial aid against the Turks from the diet, the German assembly of estates. Under the leadership of the princes, the diet reminded him that Germany's capacity for defense was weakened by the current internal anarchy. In 1455 six electors proposed to the king the establishment of an imperial court of justice in which all three estates (electors, princes, and cities) should be represented. Frederick dismissed the scheme as an attempted invasion of his authority and stubbornly maintained his disapproval in a series of stormy interviews.

In time an increasing number of princes became convinced that reform would make no significant progress until Frederick was removed. As early as 1460 the Wittelsbach princes urged his deposition in favour of George of Poděbrady, the able and resourceful king of Bohemia. To check the danger, Frederick began to dole out reforms with a sparing hand. In 1464 he consented to make the court of the treasury (*Kammergericht*) independent of his person, to staff it with representatives of the three estates, and to extend its jurisdiction into fields other than financial. It was the acquisition of Austria in 1463 on the death of his brother Albert that finally proved his undoing. The unruly Austrian nobility early took the measure of Frederick and thereafter disregarded his authority. On the east and south the duchy was imminently threatened by the expanding kingdom of Hungary under its land-hungry ruler Matthias Corvinus. The southern borders of the Habsburg lands were also ravaged by the Turks. Frederick's continuing irresolution and passivity encouraged Matthias Corvinus, who had already seized a portion of Bohemia, to launch a campaign against Austria. The Austrian nobility made no move against him, and Vienna fell to him in 1485. Frederick fled to Germany and made pitiful appeals for help to the princes. His misfortune provided the party of reform with a long-awaited opportunity. Led by Berthold of Henneberg, the able and resolute archbishop of Mainz, they pressed the aging and afflicted Frederick to relinquish the kingship in favour of his son Maximilian. Solaced somewhat by the assurance of a Habsburg succession, he gave a reluctant acquiescence, and Maximilian was elected on Feb. 16, 1486. Frederick retained the title of emperor, held since his imperial coronation at Rome in 1452. But he played no part in the government of Germany, and his death on Aug. 19, 1493, passed almost unnoticed.

Developments in the individual states to about 1500. *The princes and the Landstände.* In the various principalities the outcome of the struggle between the territorial princes and the assembly of estates (*Landstände*) was not fully decided by 1500. The vigour of the conflict arose partly out of the contrasting conceptions of government held by the protagonists. The secular princes looked upon their lands as private possessions that could be divided by agreement among their sons and drew little distinction between their private and their public revenues. The three estates regarded themselves as the corporate representatives of the whole territorial community and maintained that actions by a prince affecting their interests and privileges should be subject to their consent. They therefore opposed the partition of the territory by family pact among the princes' sons. The inadvisability of breaking up the principalities into petty territorial lordships was at length conceded by the more prudent princes. By 1500 the rulers of Bavaria, Brandenburg, Saxony, and Württemberg had accepted the principles of territorial indivisibility and primogeniture.

Dispute over taxation

In financial matters the imposition of extraordinary taxes (*Notbeden*) remained the crucial issue between the princes and the estates. The mounting cost of war and administration outstripped the ordinary revenues of the ruler, plunged him deeply into debt, and compelled him to seek financial aid from the estates with increasing frequency. In the absence of a clear distinction between public and private revenue, the estates often contended that the deficit was a private debt of the prince and disclaimed responsibility. Needy princes were thus forced to buy temporary solvency by concessions that later shackled them in their dealings with the estates. The estates regarded the *Notbede* strictly as an occasional emergency tax and insisted that it should be reasonable in amount. Indeed, the estates of Bavaria and Brunswick extracted from their respective princes in the course of the 14th century a formal recognition of their right of armed resistance to extortionate taxation. Similarly, any prince who broke his agreements with the estates was subject to the right of resistance. Thus the aims of these territorial assemblies were mixed. They sought to preserve the privileges of the three orders, to restrain the power of the prince, and to limit taxation. They were, however, also actively interested in good government, and the more enlightened rulers usually issued their ordinances only after consultation with the estates.

The princes proceeded against these powerful and often turbulent bodies with great caution. They persistently demanded that territorial assemblies convene only at the summons of the prince. They discountenanced the widespread conviction that absentees from the assembly were not bound to pay the taxes that it voted. In consequence, the peasants, who were not represented except in the Swiss cantons, Baden, Friesland, and Tirol, remained in the grasp of the princes' tax collectors. In these directions the princes had generally made notable advances by 1500, but in the vital matter of the *Notbede* they were still obliged to bargain with the estates as equals. They had

nowhere attained their ultimate objective: to transform the tax into a regular imposition voted automatically by the estates on demand.

Beyond the confines of the assembly of estates the attempts of the princes to curb their overmighty subjects aroused vigorous resistance. The noble vassals, proud and unruly, readily combined against any prince who sought to tamper with their liberties. Wise rulers deflected the nobles' energies into useful channels by employing them as stipendiaries. Hence even the most powerful princes—the Habsburgs in Austria and the Hohenzollerns in Brandenburg—proceeded circumspectly, and the difficult task of bringing the nobility to heel was far from completed in 1500. The cities of the princely territories defended their independence no less stubbornly. The princes revoked their charters, influenced municipal elections, and forbade the cities to associate in self-defense. The struggle was most intense in the north and east, where the Hohenzollern dynasty of Brandenburg emerged as the chief foe of municipal freedom. In 1442 the elector Frederick II ("Iron Tooth") crushed a federation of Brandenburg cities and deprived its leader, Berlin, of its most valued privileges. In the Franconian possessions of the dynasty, Albert Achilles of Hohenzollern waged a destructive war (1449–50) against a city league headed by Nürnberg. He suffered a resounding defeat in a pitched battle near Pillenreuth (1450). The elector John Cicero took up the battle 38 years later, when the cities of the Altmark in west Brandenburg refused to pay an excise tax on beer voted by the assembly of estates. He discomfited the cities in the ensuing "Beer War" and radically revised their constitutions to his own advantage. On the other hand, the great cities of south Germany, enriched by the Italian trade, were more than a match for the local princes: the Wittelsbach dukes of Bavaria were decisively worsted by Regensburg in 1488.

The "Beer War"

The growth of central governments. Between 1300 and 1500 the organs of central government in the territorial states became more specialized and diversified. The parent body was the advisory council (*Hofrat*) of high nobles and ecclesiastics, whom the prince consulted at his discretion. Its business was not differentiated, and there was no division of labour among the councillors. It met at the summons of the prince and did not convene at regular intervals. Its membership was not fixed, and some advisers did not attend except at special invitation. Others were regional councillors who attended the prince only when he appeared in their locality. A body so unspecialized and fluctuating was ill-adapted to cope with the increasingly complex problems of central government. Hence in the 14th and 15th centuries a professional element of "daily" or permanent councillors was introduced. They were usually legists, trained in Italy or in the newly founded universities of Prague (1364), Vienna (1365), Heidelberg (1386), Rostock (1419), and Tübingen (1477). They were well versed in Roman law, which, with its centralizing and authoritative precepts, provided a congenial climate for the growth of the powers of the territorial princes everywhere save in Saxony and Schleswig-Holstein, where the ancient customary codes were deeply rooted. Financial administration, which required specialized skills, was placed under the direction of a separate department of government, the chamber (*Hofkammer*). An inner ring of favoured advisers, the privy council (*Geheimrat*), was also instituted to counsel the prince on affairs of state. The besetting weakness of the new administrative structure was financial. Few princes followed the example of the Hohenzollern dynasty in drawing up an annual budget and requiring financial officials to submit regular accounts to the government. On the positive side, chanceries gradually created a common German language, which Luther later used to spread his message.

German society, economy, and culture in the 14th and 15th centuries. *Transformation of rural life.* Despite the impressive advance of trade and industry in the later Middle Ages, German society was still sustained chiefly by agriculture. Of an estimated population of 12 million in 1500, only 1.5 million resided in cities and towns. Agriculture exhibited strong regional differences in organization. The more recently settled areas of the north and northeast were characterized by great farms and extensive estates that produced a surplus of grain for export through the Baltic ports. The south and southwest was a region of denser population, thickly sown with small villages and the "dwarf" estates of the lesser nobility. In the northeast the great landlords, headed by the Knights of the Teutonic Order, tightened their control of the originally free tenants, denied them freedom of movement, and ultimately bound them to the soil as serfs. In the south the heavy urban demand for grain chiefly benefited the larger peasant proprietors, who sold their surplus production in the nearest town and used their gains to acquire more land. The lesser peasantry, with their smaller holdings, practiced chiefly subsistence farming, produced no surplus, and therefore failed to benefit from the buoyant urban demand. The frequent division of the patrimony among heirs often reduced it to uneconomically small fragments and encouraged an exodus to the cities. On the other hand, landless day labourers who survived the Black Death in the mid-14th century were able to command higher wages for their services.

Serfdom in the north

In south Germany the strain of transition in rural society was heightened by the policies of the landlords, lay and ecclesiastical. Confronted by labour shortages and rising costs, many landlords attempted to recoup themselves at the expense of their tenants. By means of ordinances passed in the manorial courts they denied to the peasantry their traditional right of access to commons, woods, and streams. Further, they revived their demands for the performance of obsolete labour services and enforced the collection of the extraordinary taxes on behalf of the prince. The peasants protested and appealed to custom, but their sole legal recourse was to the manorial court, where their objections were silenced or ignored. Ecclesiastical landlords were especially efficient, and peasant discontent assumed a strong anticlerical tinge and gave rise to the localized disturbances in Gotha (1391), Bregenz (1407), Rottweil (1420), and Worms (1421). Disturbances recurred with increasing frequency in the course of the 15th century on the upper Rhine, in Alsace, and in the Black Forest. In 1458 a cattle tax imposed by the archbishop of Salzburg kindled a peasant insurrection, which spread to Styria, Carinthia, and Carniola. In Alsace the malcontents adopted as the symbol of revolt the *Bundschuh,* the wooden shoe usually worn by the peasants. They also formulated a series of specific demands, which included the abolition of the hated manorial courts and the reduction of feudal dues and public taxes to a trifling annual amount. On these fundamental points there was little room for compromise, and the outbreaks were stifled by the heavy hand of established authority. But the rigours of repression added fuel to peasant discontent, which finally burst forth in the great uprising of 1524–25 (see below).

The nobility. The lesser nobility included two distinct elements. The imperial knights (*Reichsritter*) held their estates as tenants in chief of the crown. The provincial nobility (*Landesadel*) had lost direct contact with the crown and were being compelled by degrees to acknowledge the suzerainty of the local prince. The imperial knights had been extensively employed by the Hohenstaufen emperors in military and administrative capacities and were chiefly concentrated in the Hohenstaufen possessions in Swabia, Franconia, Alsace, and the Rhineland. With the extinction of the Hohenstaufen dynasty they lost their function and rewards as a nobility of service. The revenues from their small estates sank in purchasing power as prices rose. Caste prejudice prevented them from seeking an alternative role in trade or industry. Resentful of the decline in their fortunes and fiercely independent, they clung grimly to their remaining privileges: exemption from imperial taxes and the right to indulge in private war. They stubbornly resisted the persistent attempts of the princes to reduce them to subject status, and in Trier and Württemberg especially they were given valuable aid by the provincial nobles. For purposes of defense or aggression the imperial and provincial knights combined freely in powerful regional leagues, usually directed against the local princes or cities. In the course of their chronic feuds with the cities, many knights

became mere highwaymen. Many others, who had been forced to sell their estates or who were encumbered with debts, took service in Germany or Italy as mercenaries (*Soldritter*). In east Germany the knights, though equally unruly, were far more affluent. The knightly estate (*Rittergut*) was larger and produced a profitable surplus for export. The knights sat in the assembly of estates, and taxation by the prince required their consent. They were therefore well entrenched against the encroachments of princely power.

Urban life. Urban society in 15th-century Germany was concentrated in some 3,000 cities and towns. About 2,800 of the total were extremely small, with populations varying from 100 to 1,000. Of the remainder, no more than 15 cities contained more than 10,000 inhabitants. In this restricted group three were preeminent. Cologne reached its peak in the 13th century with a population of 60,000, but sank to 40,000 by 1500 following internal disputes, expulsions, and steady emigration. In 1500 Augsburg was the most populous German city, with a resident population of 50,000. Third place was held by fast-growing Nürnberg, which counted 30,000 souls. The social unity of the citizen body had been most marked in the 13th century, when the guilds joined the dominant patrician families (*Geschlechter*) in wresting the right to form an independent city council (*Stadtrat*) from the lord of the city. In the 14th century the guild masters, methodically excluded from the council by the patrician oligarchy, broke into open revolt in Speyer (1327), Strassburg (1332), Nürnberg (1348), and elsewhere. In its economic aspect the ensuing conflict embodied an attempt by the guildsmen as industrial producers to free urban industry from the tight control exercised by the merchant patriciate. By 1500 the guilds almost everywhere had gained varying degrees of representation in the city council.

The guild system

In the meantime the guilds themselves had become increasingly oligarchical and exclusive as the established masters restricted the entry of new members in order to reduce competition. The ascent of journeymen and apprentices to the rank of master was obstructed by the imposition of excessive fees, and in many guilds membership became virtually hereditary. In consequence, the journeymen began to associate in fraternities of their own to press their demands for higher wages and a shorter working day. The masters denounced the fraternities as illegal, compiled blacklists of leading agitators, and formed intercity associations to enforce low wage rates. The day labourers and casual workers outside the guild structure had no protective organization and suffered heavily in periods of economic depression. The surviving tax records of the German cities, though not wholly reliable guides, nevertheless suggest wide extremes of wealth and poverty. In late 15th-century Augsburg, 2,985 of a total of 4,485 households (66 percent) were recorded on the tax rolls as exempt from taxation on the ground of insufficient means. At the other extreme stood the enterprising and prosperous business dynasties of the Fugger and the Welser. Not all wealthy citizens lacked public spirit, however, and hospitals, almshouses, and charitable foundations multiplied within the city walls. The spreading problem of mendicancy was combated by stringent legislation against able-bodied beggars in Esslingen (1384), Brunswick (1400), Vienna (1442), Cologne (1446), and Nürnberg (1447).

The decline of the church. The vigour and assertiveness of secular society in Germany was exercised increasingly at the expense of the clergy and the church. Among the upper clergy more than 100 archbishops, bishops, and abbots were temporal rulers. The prelates were usually sons of the nobility and did not allow election to church office to interfere with their aristocratic ardour for war and territorial acquisition. They were expert in the accumulation of benefices and were notoriously lax in the performance of their spiritual duties. Their influence was freely used to advance their kinsmen and partisans among the greater and lesser nobles, who dominated the cathedral chapters and ruled the abbeys. The monasteries were filled with monks and nuns who were distinguishable from the lay aristocracy only by a nominal celibacy. Among the secular princes, the rulers of Austria, Brandenburg, and Saxony wrested a right of appointment to a fixed number of bishoprics and lesser church offices from the papacy, which had been gravely weakened by the schism and the conciliar movement of the 15th century. All lay and ecclesiastical princes imposed heavy extraordinary taxes on the clergy. The steady invasion of the church by secular interests was also exemplified by the moral and material condition of the lower clergy. The Black Death of 1348–49 had decimated the ranks of the more dedicated priests, who ministered to their plague-stricken flocks instead of seeking safety by flight. The new recruits who rushed into holy orders were often self-seeking and spiritually unqualified. As the inflow continued, the problem of clerical unemployment and inadequate stipends attained greater proportions. Many were compelled by need to accept ill-paid livings. Others obtained no benefice at all and lived precariously as chantry priests or as itinerant chaplains. Their moral and intellectual defects were bitterly assailed by church reformers and by an increasingly well-informed laity. Many pious Christians, especially in the cities, began to turn away from the priesthood in their search for spiritual comfort and to seek relief in mysticism or in lay associations practicing a simple, undogmatic form of Christianity.

Trade and industry. The most impressive achievements of the German economy between 1200 and 1500 lay in trade and industry. German trade benefited from the Hundred Years' War between France and England, which diverted northbound Mediterranean merchandise from the customary Rhône valley route to the eastern Alpine passes; from the fierce internal warfare between the Italian city-states, which weakened their supremacy in long-distance trade; and from the rapid economic development of "colonial" eastern Europe between the Baltic and the Danube. The north German trade was chiefly based on staple commodities such as grain, fish, salt, and metals; but the south German merchants, in their capacity as middlemen between Italy and the rest of Europe, had taken the lead by 1500. They combined trade and industry in the great Ravensburg Trading Company (1380–1530), which produced and exported Swabian linen and laid the foundation of the fortunes of the Höchstetter, Herwart, Adler, Tucher, and Imhof families. The most important independent concern was that of the Fugger, whose founder, Hans Fugger, began his career as a linen weaver in Augsburg. The Fuggers' accumulated profits provided capital for moneylending and banking, which they conducted with the aid of business techniques borrowed from the more advanced Italians. The wealth and prosperity of Germany, which Machiavelli remarked on in 1512, stood in sharp contrast to its political and military weakness, a disparity that contributed significantly to a profound sense of malaise and discontent on the eve of the Reformation.

The Ravensburg Trading Company

Cultural life. In the absence of a strong centralized monarchy to act as a focus, German culture continued to be regional in character and widely diffused. The mysticism of Meister Eckhart, Johann Tauler, and Heinrich Suso, which commanded all men to look for the kingdom of God within themselves, flourished chiefly in the cities of the Rhineland, where lack of diligent pastoral care forced Christians to call upon their own inner resources. In the same region social and moral satire attained an urgent and vivid realism. Sebastian Brant (1458–1521), born at Strassburg, spared no class in his epic on human stupidity, the *Narrenschiff*, or *Ship of Fools*. But it was in the thriving cities of south Germany, as yet little affected by Italian humanism, that late Gothic culture reached magnificent heights in art, architecture, and sculpture. Albrecht Dürer, born in Nürnberg in 1471, challenged his generation with his evocative engraving of "Melancolia I" in which a brooding figure with closed wings sits idly amid a chaos of scientific instruments and meditates on the futility of human endeavour. In architecture the hierarchical elaboration of the late Gothic style maintained its ascendancy and even made a notable conquest in Italy with the construction of the great cathedral of Milan, begun in 1387. The sculptured carvings of Tilman Riemenschneider (*c.* 1460–1531) in the castle of Würzburg revealed the anxiety, the deep piety, and the religious sensibility of

Christian men engaged upon a spiritual pilgrimage that was to continue to the Reformation and beyond. His work was the pinnacle of a great flowering of sculpture, one of the greatest in German history. (C.C.B./L.G.D.)

Germany from 1493 to *c.* 1760

REFORM AND REFORMATION: 1493–1555

The empire in 1493. The reign of Maximilian I (1493–1519) was dominated by the interplay of three issues of decisive importance to the future of the Holy Roman Empire: the rise of the Austrian House of Habsburg to international prominence, the urgent need to reform the empire's governing institutions, and the beginnings of the religious and social movement known as the Reformation. The accession of the dynamic and imaginative Maximilian to the German throne aroused in many Germans, and in particular among humanists, expectations of a time when the old imperial idea—the vision of the empire as the political expression of a united Christendom in which the emperor, as God's deputy, rules over a universal realm of peace and order—might become a reality. Since the extinction of the Hohenstaufen dynasty in 1254, imperial authority had been in disarray; as weak emperors had become absorbed in struggles against foreign and domestic, secular and ecclesiastical rivals, real power in the empire had moved toward the governments of territorial states and independent cities. From their first appearance on the historical scene (briefly from 1273 to 1308, then from 1438 in a nearly unbroken line until the dissolution of the empire in 1806), Habsburg rulers had fostered imperial unity. But they had been notably unsuccessful in creating agencies for its attainment, partly because they were assiduously working to build up a power base for their own house. This, in turn, brought them into conflict with European antagonists, chiefly France. The long reign of Maximilian's father, Frederick III (1440–93), was regarded by nationalists as lamentable in its inattention to the problems pressing on the empire. The solutions proposed for them were subsumed under the name of "reform," a highly charged word that acquired enormous additional force in the 15th century when the conciliar movement and its lay and clerical proponents exerted pressure for religious renewal. Maximilian's arrival on the throne thus generated a surge of anticipation, expressed in an outpouring of agendas for restructuring what was then coming to be called the "Holy Roman Empire of the German Nation."

Imperial reform. Inevitably, perhaps, Maximilian's performance with regard to the empire disappointed. But he was successful in significantly extending the dynastic might of his family. He took over the duchy of Tirol with its vast mining resources. Ambitious marriage alliances spread Habsburg entitlements west and east: in 1496 Maximilian's son Philip wed Joan, the daughter of the king and queen of Spain, thus linking Habsburg Austria to Spain and the Netherlands (the future Charles V was born of this union in 1500); and in 1516 Maximilian's grandson Ferdinand was betrothed to the heiress of Hungary and Bohemia. But these connections only escalated Maximilian's internal and external problems. In foreign politics his ventures ended, for the most part, in calamities for the empire. Switzerland was lost (1499), several Italian campaigns were repulsed, and an attempt against the duchy of Burgundy brought the hostility of France and led to the fall of Milan. Even the imperial crown eluded Maximilian: his advance on Rome for this purpose was halted, and he had to be content with the self-bestowed title of "Roman Emperor Elect."

These reverses strengthened a reform party among leading members of the empire's estates (Stände), especially their spokesman, the archbishop-elector of Mainz, Berthold von Henneberg. Given the long rivalry between emperor and estates, it goes without saying that their respective plans for reforming the empire diverged on crucial points of direction and control. The estates, acting from their—to them entirely legitimate—sense of the prerogatives of particularism, favoured a central administration responsive to them; the emperor, to the contrary, insisted on organs subservient to him, modeled on bureaucratic agencies recently established in Burgundy and Austria and shored up by the authority-enhancing principles of Roman law.

At the Imperial Diet held in the city of Worms in 1495 the estates, whose members had begun to see themselves as the authentic representatives of the whole country, prevailed. The four reform measures adopted on this occasion were in large part intended to limit the emperor's powers. An "Eternal Peace" outlawed private feuds, and steps were taken, agreed to by the emperor, to implement this pacification. An Imperial Chamber Court functioned as supreme tribunal for the empire, most of its judges being named by the estates. An empirewide tax, the "Common Penny," was imposed, the collection of which fell to the estates. To these measures was added, in 1500, an Imperial Governing Council to monitor, under the domination of the electors, the empire's foreign policy. Maximilian was able to impede the operation of this body (it was reestablished in 1521), but overall the emperor's reform objectives had come to nothing.

The historical judgment on this failure has swung between scorn for Maximilian's imperial "fantasies"—for whose realization a reorganized Germany was to furnish the means—and respect for his grasp of the country's perilous geopolitical situation, and between sympathy for the estates' single-minded pursuit of the imperatives of regional state building and disdain for the parochialism of their political vision. Most historians have found Maximilian's actions distorted by a species of romantic idealism (he in all seriousness proposed to Bayzid II that Turks and Christians should settle their differences by the rules of the tournament; he also cherished for a time the hope of becoming pope as well as emperor). What is clear is that the reform effort so briskly launched at the beginning of his reign resulted in permanent confrontation between sovereign and estates, a posture in which neither party could outmaneuver the other and every matter of policy required tenacious negotiation. In light of this permanent tug of war, the notion of a universal realm ruled by a "German" emperor does seem fantastic, though it never lost its ideological force. No one in the empire possessed plenary authority, but real power was shifting to territorial rulers and their bureaucratic agents and to the magistrates of larger cities. This shift was to be of the greatest importance in determining the direction in which the Reformation established itself in Germany.

The Reformation. The Reformation presents the historian with an acute instance of the general problem of scholarly interpretation—namely, whether events are shaped primarily by individuals or by the net of historical circumstances enmeshing them. The phenomenon that became the Protestant Reformation is unthinkable without the sense of mission and compelling personality of Martin Luther. But in social and intellectual conditions less conducive to drastic change, Luther's voice would have gone unheard and his actions been forgotten. Among the preconditions—which are the deeper causes of the Reformation—the following stand out: (1) Everyone agreed that the Roman Catholic church was in need of correction. The lack of spirituality in high places, the blatant fiscalism, of which the unrestrained hawking of indulgences—the actual trigger of the Reformation—was a galling example, and the embroilment in political affairs were symptoms of corruption long overdue for purgation. While the church continued to be accepted as the only legitimate mediator of divine grace, denunciations of its abuses, perceived or actual, became more strident in the decades before 1517. (2) A subtle change, moreover, had been occurring in people's religious needs and expectations, leading to demands for a more personal experience of the divine. Failing to meet this aspiration, the church was widely, if diffusely, rebuked for its unresponsiveness. (3) More focused criticism came from the Christian humanists, an influential group of scholars bent on restoring the fundamental texts of Western Christianity. Led by Desiderius Erasmus, the most renowned biblical scholar of the time, these men held the Catholic church up to the spiritual ideals for which it claimed to stand and, finding it wanting, set the principle of Evangelicalism against the church's secular-

Causes of the Reformation

ized ambitions. (4) Still more fatefully, by 1500 the church had come under attack from European rulers whose administrative, legal, and financial hegemony could not be completed in their respective states without domination of the ecclesiastical sector. In the empire, as elsewhere, the trend in ecclesiastical politics was toward state churches (*Landeskirchen*), in which governments, using "reform" as a pretext, gradually gained, while church authorities lost, a large measure of control over clerical properties, personnel, and functions. The Reformation was the culmination of this process, which, in the empire, took place in nearly all princely territories and in most independent cities where governments brought the administration of the church under political direction. (5) In Germany this development was facilitated by an ancient feudal custom entitling a landlord to extend "protection" to churches located on his estates. Over this "owner's church" (*Eigenkirche*) he enjoyed the right of patronage, allowing him to appoint incumbents and manage properties. In the course of extending their sovereignty, territorial princes took over this right to patronage and fashioned of it the legal basis on which, in the Reformation, they assumed full control over the administration of the church. (6) In every segment of German society, but particularly among the poor, voices were being raised against injustice and exploitation. Wide disparities in income and discriminatory laws in cities as well as the deteriorating standard of living of small peasants and agricultural labourers caused riots and uprisings, which by the early 1500s had become endemic.

These, then, were the forces driving events toward a crisis. In the first decade of the 16th century they coalesced into a powerful surge of religious, social, and political agitation, for which "reform" (of church and society) was the code word. Ironically, Luther, who was to channel this agitation into the Reformation, had, until his emergence as a national figure in the 1520s, nothing to do with it. For him one issue alone mattered: the imperative of faith. His personal path to the Reformation was an inner search for religious truth, to which his conscience was his guide.

Martin Luther

When he wrote his Ninety-five Theses against indulgences in October 1517, Luther was an Augustinian friar, a preacher in the Saxon city of Wittenberg, and a theology professor at the university founded there in 1502 by the elector of Saxony, Frederick III, called "the Wise." His ambitious father had pushed him toward a career in law, but in 1505 the fervently devout Martin entered a monastic house. His order, that of the Augustinian eremites, was a strict reform congregation dedicated to prayer, study, and the ascetic life. Deeply troubled by the question of justification—of how a human being, a sinner, may be justified (saved) in God's sight—Luther found no comfort in monastic routine and turned to an exploration of the sources of the Christian faith, notably St. Paul and St. Augustine. His intellectual promise having been recognized, he was sent by his order to study theology at Erfurt and Wittenberg. He gained his doctorate in 1512 and commenced his teaching of the Bible in Wittenberg that same year. According to his own account, it was during his close reading of Paul's Epistle to the Romans, while preparing to give a course of lectures on that text, that he discovered what struck him as the solution to the problem posed by the huge gap between human sin and divine grace. Justification is not earned as a reward for human effort through good works (a position Luther now attributed to a misguided and misguiding Roman church). To the contrary, human beings are justified without any merit of their own by God's freely given and prevenient (*i.e.*, coming before any worthy human deeds) grace, through faith, which is a gift of God. This is the meaning Luther found in the crucial passage in Romans 1:17: "For in it [*i.e.*, the gospel] the righteousness of God is revealed through faith for faith: as it is written, 'He who through faith is righteous shall live.' " "Righteousness"—*justitia* in Latin—does not refer, Luther now saw, to God's activity as judge but to the justifying righteous condition he effects in the human sinner, a condition expressing itself as faith. The momentous consequences of this theological insight, which Luther appears to have taken as a unique discovery but which had in fact been espoused by a score of theologians before him, were not then apparent to him. They asserted themselves powerfully, however, once he began to lecture and preach on the—for him—paramount themes of salvation by faith alone (*sola fide*) and exclusive reliance on Scripture (*sola scriptura*). It was the indulgence controversy of October 1517 that brought it all into the open.

Few other issues could so clearly have exposed the gulf that separated this ardent friar from an urbane and pragmatic church. The indulgence offered in Saxony in 1517 had its origin in two purely financial arrangements. First, Popes Julius II and Leo X needed funds for rebuilding Saint Peter's Basilica in Rome; second, Bishop Albert of Hohenzollern, forced to buy papal dispensations in order to gain the archbishoprics of Mainz and Halberstadt, agreed to promote indulgences in his domains, half the income from which was to go to Rome, the other half to him and his bankers. For Luther, the issue turned not so much on the outrageous venality of this deal as on the indulgence itself. Truly contrite sinners do not desire relief through an indulgence (which is a remission of temporal punishment to be performed following absolution); they crave penance. This is the gist of Luther's argument in the Ninety-five Theses, which he sent to his ecclesiastical superiors to persuade them to abandon the indulgence sale. (The story that he nailed a copy of the theses to the door of the castle church in Wittenberg may be the invention of a later time.)

Luther intended no defiance with this action. He intervened as a priest on behalf of his flock and as a conscientious theologian against a corrupting church. But the public reaction to the theses (he had written them in Latin, but they were soon translated and printed) made it evident that he had touched a nerve. Encouraged by expressions of support and goaded by opponents, Luther became more outspoken, harsher in his criticism of the church, and more focused in his attacks on the papacy. By 1520 he was well on his way to becoming the spokesman for Germany's grievances against Rome. A pamphlet he published that year, "Address to the Christian Nobility of the German Nation," urged the empire's secular rulers to reform a church that would not set its own house in order. Popes and prelates are not sacrosanct, he argued; they may be brought to justice. As every Christian can read the Bible for himself, papal claims to interpretive authority are a vain boast. Luther prodded the German princes to consider the state of the church and to reform it for the sake of the faith. In this way Luther drew out, albeit reluctantly, the full consequences of his principle of "salvation by faith alone." No church was needed to act as God's agent; grace was available without mediation. No priest, not even the pope, has special powers, for, so Luther argued, all human beings are priests, made so by their faith. It is scarcely surprising that a bull of excommunication against him (*Exsurge domine*) issued from Rome in June 1520.

Imperial election of 1519 and the Diet of Worms. At any other time the Lutheran matter would probably have ended there. But 1521 was no ordinary moment in the empire's history. When he died in 1519, Maximilian had not succeeded in having his grandson and heir, Charles, designated his successor. King of Castile and Aragon since 1516, and suzerain also over Habsburg apanages in the Netherlands, Naples, and central and eastern Europe—not to mention the Spanish possessions in the New World—Charles posed a formidable problem for the electors. Should they choose a man whose vast resources might well empower him to centralize authority in the empire when this would limit their hard-won autonomy? Should they elect a prince whose international commitments would entangle Germany in European conflicts? The only other viable candidate was the king of France, Francis I, who recognized in the rapidly growing Habsburg ascendancy a serious threat to his own power. But, though Francis made tempting promises, the electors, in the end, opted for Charles—not, however, without drawing from him a 36-point "electoral capitulation" in which he swore to uphold the estates' prerogatives. While in so choosing the electors strove to secure their own privileges, the German public tended to see in the new monarch a fulfillment of national

aspirations—mistakenly, as it turned out, but giving rise, at the time, to patriotic expressions of hopes for a brighter future. Coinciding with these events in the political realm and gaining resonance from them, the Lutheran matter was inevitably drawn into the political vortex, thereby acquiring the features of a national cause.

The Diet of Worms of 1521

Delayed by disturbances in his Spanish domains, Charles reached Germany late in 1520. He was crowned king in Aachen, assuming at the same time the title of Roman emperor elect. He then proceeded to Worms, where he was to meet with the German estates in early 1521. By then no other issue counted as much on the agenda as the Lutheran affair. Acting out of what appears to be a blend of conviction and political expediency, the estates' leaders, prompted by Frederick the Wise, demanded that the diet reopen Luther's case by allowing the excommunicated friar to speak before the estates. Unable to resist, the emperor issued a safe-conduct, and Luther traveled to Worms, treated on his way to so extravagant a public acclaim that he must have felt like a national hero. On April 17 and 18 he stood before the emperor and representatives of the estates. He refused to revoke his views on justification and other points of theology, reiterated his denial of ultimate authority to pope and church councils, and—when pressed—asserted the principle of individual responsibility in matters of faith: "As long as my conscience is held captive by the words of God, I cannot and will not revoke anything, for it is dangerous, and a great peril to salvation, to act against conscience. God help me. Amen." These words, and all else that transpired, were broadcast to the country in scores of pamphlets in which Luther was cast in the role of the man sent by God to cleanse the German church. While an imperial edict condemned his teachings and placed him and his adherents under the ban, Luther himself was offered a refuge in Frederick the Wise's castle in the Thuringian forests, the Wartburg. There, fretting over his enforced absence from events, he turned to the translation of the New Testament, while the movement, of which he was now the acknowledged head, gained momentum.

Analysis of what happened at Worms reveals the first signs of what was to become a fateful split in the self-perception of this movement. The estates saw it as a means of promoting ecclesiastical reform; on questions of faith, they were willing to compromise. Luther, on the contrary, tolerated no distinction between points of faith and aspects of church practice, and on his understanding of the former he stood rock solid. He and his political patrons were thus pursuing different ends; for the present, however, his support in official circles was, for whatever reason, substantial.

Luther's hold on the general public was even more impressive than his hold on the political leadership. Historians are not unanimous in explaining how this friar from the German east, utterly obscure only five years earlier, could have gained such a following by 1521 that few governments would enforce the ban against him, knowing that they would face strong resistance if they tried to do so. It is not clear whether the general population, still largely illiterate, understood the doctrine of solifidianism well enough to make it the object of informed choice. More likely, the German populace took Luther not as the preeminently religious prophet he was but saw in him their best hope of achieving an amelioration of the many troubles vexing them in their respective stations, not only in religion but also, and perhaps mainly, in their social condition. Against these conditions, the products of deeply rooted legal and institutional structures, Luther—or so people seem to have understood him—raised the standard of evangelical morality: all things were to be judged by Scripture and God's law.

The issue of social reform

Scholars have often pointed out that this view is less attributable to the Wittenberg reformer than to another theologian and preacher then beginning to be active in the Swiss city of Zürich, Huldrych Zwingli. Luther explicitly rejected the use of the—to him purely spiritual—New Testament as a norm for social reform. But Zwingli affirmed it. On the other hand, Luther's vehemence and forceful rhetoric were so compelling that, rightly or wrongly, his name came to be fused with the general hope of improvement in human affairs. His partisans, proficient in their use of the media of print and woodcut illustration, helped shape this conviction by furnishing propaganda for a strong popular drive toward Lutheranism. This drive from below, further nourished by traditional anticlerical sentiments, was met from above by the eagerness of territorial and urban governments to utilize Lutheran ideas as legitimation for the extension of political control over the church. Thus, by the mid-1520s, a number of German cities and states had formally turned Lutheran, meaning that they had severed their legal and administrative ties to Rome and its prelates and were building new ecclesiastical institutions and framing new doctrines.

The revolution of 1525. The events of the revolutionary years 1524–25 underscored the urgent need to establish Lutheran church organizations. In its own time, and often since then, these events were labeled a "peasant rebellion"; but modern scholarship has made it clear that the insurrection was far more than a series of uprisings by rural bands. The tens of thousands of peasants drawn into the movement, some of them massed in major military actions, were a symptom of the general unrest that had gripped Germany since the middle of the 15th century. This unrest was ultimately caused by demographic pressures with attendant economic and political dislocations. The particular demands pressed in the 1520s—mitigation of fiscal and labour burdens imposed on peasants by their lords, autonomy for village communes, and relief from high taxes—had been voiced before. New was the linkage of these demands with the grievances of restive urban groups also protesting exploitation and disenfranchisement and their formulation as an agenda of social reform on the principle of Christian communitarianism. This ideological redirection of old patterns of resistance could not have occurred without the impetus of the Reformation, specifically the incendiary preaching in towns and villages of evangelical pastors who represented Lutheran and Zwinglian ideas as solutions to the problems at hand. The clearest evidence of the Reformation's impact on the shaping of what some modern scholars call "the revolution of the common man" are the Twelve Articles drawn up for the Swabian peasantry by an evangelical cleric and associate of Zwingli's. Article three of this document asserts that "The Bible proves that we are free, and we want to be free," while another article claims for every congregation the right to choose its minister; these articles are a strong indication of how vital the principle of Biblicism had become to people preeminently concerned with worldly life.

In the regions involved—Franconia, Swabia, the Upper Rhine and Alsace, Thuringia, and Tirol—large forces of peasants attacked castles, monasteries, and some cities. News of these actions encouraged discontented urban groups to rise against their oligarchic town governments, and for a while it looked as though a united revolutionary front of ordinary—*i.e.,* nonprivileged—people might be forged. Manifestos and lists of articles abounded; there was talk everywhere of judging things by "God's law" (meaning the gospel), and some groups even laid plans for a "Christian association" across regional and town–country lines. But before long the forces of authority won the upper hand, and the insurrectionaries were put down with the ferocity customary in those days. The war's final stage was dominated by Thomas Müntzer, a visionary theologian with a message of social deliverance for and by the poor. The defeat of his forces at Frankenhausen in May 1525 marks the final victory of the old order over the would-be new dispensation.

Thomas Müntzer

Luther was heavily implicated in this turnabout. Realizing that his words and deeds had served to encourage popular action against legitimate rulers, he sought to separate himself drastically from the movement, going so far as to urge the rulers' warriors to "cut [the peasants] down, hit them, choke them wherever you can." His brush with revolution confirmed Luther in the rigorous segregation he made between "two realms," the worldly and the spiritual, the respective laws and ideals of which must not be confounded, as, he claimed, the revolutionaries had done. The gospel, he said, cannot be used as a standard

for governing in the world, which has its own rules and ways of justice, many of them, he acknowledged, unfair and blameworthy. Luther's separation of worldly and evangelical values, soon made binding law by Lutheran governments, brought an abrupt end to the early phase of the Reformation, during which events seemed to many to be moving toward a sweeping transformation of social, as well as religious, structures. At the same time, Lutheran governments read 1525 as a lesson on the need to control subjects, concluding that preaching in particular, and religious behaviour generally, must be controlled. To accomplish this purpose, new laws and bureaucracies were set up everywhere.

Lutheran church organization and confessionalization. The 1525 revolution was but one of several upheavals worrying the authorities. Three years earlier a group of imperial knights led by Franz von Sickingen had declared feud on the archbishop of Trier, claiming to derive from Scripture their right to despoil Roman Catholic prelates. The ensuing "Knights' War" was quickly crushed. But about the same time a disturbance broke out in Wittenberg where, during Luther's exile in the Wartburg, a group of reforming Spiritualist activists forced the city council to abolish many traditional Catholic practices. Upset by this rash move, Luther intervened to reverse it. But this incident, and the knights' attack, caused consternation among the heads of government, who feared loss of control. Their anxiety was deepened by the spread from Switzerland in the mid-1520s of Anabaptism, a radical religious movement whose most distinctive tenet was adult baptism. The events of 1525 thus strengthened a long-ripening resolution that firm structures and clear doctrines were needed to reassert authority in a situation of drift.

The ability of Lutheran states to act on this resolution was facilitated by the impact of foreign affairs on the empire's internal politics. Far from seeing to the execution of the 1521 edict against Luther, Charles V left his brother Ferdinand in charge of imperial affairs and departed from Germany after the Worms diet to deal with the many problems besetting his far-flung interests. The most perilous of these was the war with France, which implicated the emperor in a constantly shifting balance of alliances with other powers and in a see-saw of military actions in which now he, now Francis I, was in the ascendant. Charles's victory at Pavia in 1525 led in turn to the formation of a coalition against him (the so-called "Holy League of Cognac"), intended to forestall Habsburg hegemony in Europe (a scenario to be replayed many times in the following two centuries). In 1526, therefore, Charles was in no position to dictate to the German estates on the Lutheran matter. Within a year, however, the situation turned in his favour. Spanish troops captured and plundered Rome in 1527, and by 1529 Charles was dominant once more, though it had become clear that neither warring party could bring the other to its knees.

The Turkish advance

At the same time a potentially fatal danger loomed in the east where the Turks, under Süleyman I the Magnificent, began to aim their path of conquest at the Balkans and Hungary. The death of King Louis II at the Battle of Mohács in 1526 put Ferdinand of Habsburg (Charles V's brother) in line for the Bohemian and Hungarian crowns, thereby exposing the already overextended Habsburgs on a new front. By 1529 the Turks were moving toward Buda (captured in September of that year) and Vienna. Facing these perils, Charles concluded peace with France, sealing his triumph in the west with his coronation as emperor at Bologna. He then returned to Germany.

The events just described formed the larger political frame in which Lutheran church organization took place. Forced to solicit military aid from the estates in 1526, Ferdinand postponed implementation of the Worms edict, accepting a declaration by the Diet of Speyer of that year to the effect that every estate "will, with its subjects, act, live, and govern in matters touching the Worms edict in a way each can justify before God and his Imperial Majesty." This declaration gave Lutheran rulers the signal to proceed with their intended legal, administrative, financial, and liturgical reforms, and the years following 1526 saw the construction in every Lutheran territory of what amounted to a state church, headed by the ruling prince. In 1529 this process was interrupted when, following the emperor's military successes, Ferdinand demanded at a diet, also held in Speyer, that, pending a general council to decide the religious issue, Lutherans should end their separation. (It was the "protest" of a number of princes and cities against this abrogation of the earlier Speyer decree that attached to Lutherans the name "Protestants.") By then, Protestants were no longer a united party. Luther and Zwingli had met at Marburg in 1529 in an attempt to iron out differences, but they could not agree on the question of Christ's real presence in the Eucharist. While a few Lutheran princes prepared for military action, a compromise-minded group led by the humanist Philipp Melanchthon, who dreaded the prospect of fragmentation within Protestantism, drew up a moderate outline of Lutheran positions. These were presented for discussion at the Diet of Augsburg in 1530, which was attended by the emperor. The Augsburg Confession, which became a fundamental statement of Lutheran belief, assumed that reconciliation with the Catholics was still possible. This view was shared by Charles, who was pushing the pope toward the summoning of a general council to mend the religious split. Negotiations among theologians and politicians, however, came to nothing, and the end result was that, with the Augsburg Confession rejected, Lutheranism was outlawed again. The militant Protestant faction, led by Philip, landgrave of Hesse, now established a formal organization of resistance, the Schmalkaldic League (1531), and the empire moved toward armed conflict as Lutherans passed beyond the formation of a political party and became a military force as well.

Around this time much more rigid standards of religious orthodoxy and conformity were imposed. This development has been called "confessionalization," a concept used by some historians to define developments in the empire during the second half of the 16th century. Confessionalization completed the process, under way since the late Middle Ages, of meshing religious and church politics with the objectives of the state. Central to this process was the institution of a territorial religion that was based on an authorized declaration of doctrines (a "Confession") binding on all subjects and implemented by an established church responsible to the ruler (or, in city states, to the magistrates). Tending toward exclusiveness and therefore intolerance, this system contributed to the warlike turn taken by events after 1530. More important in the long run, confessionalism promoted a social drive, also long under way, toward the inculcation of discipline and order in public and private as well as in religious and civic affairs. Through catechisms, schooling, family and welfare legislation, norms for work, and standards for personal life, state and church attempted to restructure society in accordance with the goals of what has since been called the Protestant temperament. Success was slow in coming and never more than partial. But there is no doubt that, through the confessional process, Protestantism left a deep imprint on the German character.

Confessionalization

Religious war and the Peace of Augsburg. After the diet of 1530 Charles left Germany for more than a decade, occupied with troubles in the Mediterranean, the Netherlands, and, once again, France. In 1535 he campaigned against Tunis to subdue the Barbary pirates who, as the naval arm of the Ottomans and as corsairs and privateers, had been making navigation unsafe. Renewed war with France was temporarily halted in 1538 by a treaty meant to last 10 years, but in 1542 France struck again, allied now with major European powers, including the duke of Guelders and Cleves, whose lands were claimed by Charles as part of his Burgundian inheritance. The emperor's conquest of this duchy in 1543, which considerably broadened his power base, and the peace he concluded with France in 1544 (Peace of Crépy), followed by an armistice in 1545 with the Ottoman Empire, left him free at last to deal decisively with the German Protestants.

The emperor's policy toward religious deviants was guided by his concept of empire. The universal realm over which he hoped to reign faced external and internal threats; its mission of unity and order was assaulted by

infidels from without and by national rivalries and heresy from within. He had dealt with the first and second threats; now he turned his attention to the third. Protestantism had spread rapidly in Germany. More than a religion, it was, by the 1540s, a full-fledged political movement with a growing military capacity. The number of Protestant territories had recently grown to include, among others, Brandenburg, the Palatinate, Albertine Saxony, and the bishoprics of Cologne, Münster, Osnabrück, Naumburg, and Merseburg. In Philip of Hesse the Lutherans had an able political strategist. At least provisionally, pending the settlement of all religious issues by a general council, the Protestants had won grudging recognition of their right to exist. Such a council was actually summoned by Pope Paul III—though only upon repeated prodding by the emperor—but there were few signs that the Protestant states would submit. In 1545, therefore, Charles decided on war. He found a pretext in the capture, by Lutheran princes, of the duke of Braunschweig-Wolfenbüttel, a Catholic who had tried to reconquer the lands from which he had been expelled by his Lutheran subjects. Claiming that this capture violated imperial law, Charles opened the conflict in 1546, in which he was joined by Maurice, duke of Saxony, an ambitious Lutheran prince to whom Charles had secretly promised the Saxon electorship. The ensuing war fell into two phases, the first of which saw the emperor victorious (Battle of Mühlberg, 1547). Capitalizing on this strong position, Charles in 1548 forced the estates to accept an Interim, a temporary religious settlement on the emperor's terms. It was the political concessions Charles demanded from the estates, however—concessions that would have permanently limited their autonomy—that led to a resumption of war. Among the Protestants the lead was now taken by Maurice of Saxony, who had abandoned the emperor and had obtained material support from the new French king, Henry II, for fighting on the Protestant side. The resulting "Princes' War" was brief (1552–53) and inconclusive, and in 1555 a peace was signed at an imperial diet held, again, in Augsburg.

The Peace of Augsburg

The Peace of Augsburg closed one epoch of German history and opened another. It decided the religious issue but did so in a way bound to occasion future problems. It reinforced the princes' authority over their territories but failed to settle their relations with the emperor. Most important, it legalized Lutheranism, laying down the rule, later epitomized in the phrase *cuius regio, eius religio* ("he who governs the territory decides its religion"), that each estate—*i.e.*, each prince or city—could opt for either the Roman Catholic or the Lutheran religion (*jus reformandi*) and that this choice was binding on everyone under that ruler's jurisdiction. Only one faith could legitimately exist in a given state, and that faith had to be the ruler's and could be only Catholicism or Lutheranism; Calvinism, Zwinglianism, and Anabaptism were excluded. A subject unwilling to live by this choice was free to emigrate and take his belongings with him (a provision considered liberal at the time). Confiscated church properties could be kept by the governments that had taken them. An Ecclesiastical Reservation prevented ruling prelates from converting their lands along with them. These terms make it clear that the real winners of the war, and of the entire Reformation period, were the territorial princes, whose authority and power, which now encompassed the church, were greatly increased. It scarcely needs to be said that Luther, who had died in 1546, would not have approved of this outcome. As for the emperor, he abdicated in frustration and retired to a monastery in Spain, leaving his Spanish and Burgundian crowns to his son Philip and the empire to his brother Ferdinand. These two men, as Philip II and Ferdinand I, strong-minded Catholics both, were to play prominent roles in the period of Counter-Reformation and confessionalism that dominated Europe after 1555.

THE CONFESSIONAL AGE: 1555–1648

German society in the later 1500s. The changes caused by state building and the Reformation bestowed little real benefit on the lives of ordinary people. Historians agree that the later 16th century was, for many, a time of economic hardship and social stress. Rapid increase in population (the European population rose by more than half between 1500 and 1700) and, secondarily, the influx of precious metals from the New World were the main causes of an inflationary trend that spanned the entire century and reached painful stages in Germany in the 1590s and early 1600s. Grain prices were especially affected, with the result that an ever smaller share of the ordinary person's budget was available for the purchase of other products. This had several effects, which, at least in outline, are well documented. The quality of nutrition for all but the wealthiest became much worse than it had been in the late Middle Ages, when meat consumption was at an all-time high. Illness and epidemic disease were frequent as the nutritional deficiency was aggravated by a series of bad harvests, perhaps caused by unusually severe winters in the decades after 1560.

Cities and towns suffered loss of income as the market for their manufactured wares declined. In consequence, municipal guilds lost ground, not only economically but also politically, owing to the curtailment of their participation in urban policy-making. There were exceptions to this trend. Craftsmen specializing in the manufacture of luxury cloths and arms found lucrative markets at princely courts; but overall the position of artisans declined. Journeymen could no longer anticipate becoming masters. Artisans employed in traditional handiwork felt the pressure of the putting-out system favoured by early capitalism, whereby much production was moved from the town to the countryside. Division of labour was introduced, gradually transforming self-employed craftsmen into dependent workers.

In the agricultural sector, high grain prices and rising land values improved the lot of peasant proprietors, but the real beneficiaries were landowning nobles and urban patricians with investments in agriculture. Society was polarized by these developments. A minority of rich peasants lived amid struggling smallholders hard-pressed by landlords who maximized their profits by increasing labour and tax burdens (the period has been spoken of as that of a "second serfdom"), and in the cities an upper crust of merchants and landed aristocrats faced a proletariat, whole sections of which were pauperized by the end of the century. The populous cities, once the glory of Germany, began to play a smaller role, as economic troubles and the centralizing policies of territorial princes decreased their prosperity and sapped their political strength.

Social polarization

The highly visible contrast between rich and poor and the animosity of the weak against the wielders of influence created tensions among groups and classes. Political and economic power was more concentrated than ever before. Its new centres in Germany were the splendid courts of secular and ecclesiastical princes, whence it was distributed to favoured groups: the nobility, rising in importance again but finding its function limited to the service of rulers, and the upper bourgeoisie, shifting its loyalty from guild hall to palace. For ordinary people, administrative centralization and politically sanctioned Reformation had the effect of making their lives more rigid. A host of mandates flowed from centres of government, seeking to promote an ethic of order, productivity, and morality by shaping working and domestic activities as well as private habits and attitudes. These inroads caused resentment, and there is evidence of widespread resistance, most of it passive. Under these circumstances, the evangelical Reformation seems to have made but slight impact on the populace at large, whose effective religion continued to be a mixture of traditional Christianity and folk magic.

Most people were worse off near the end of the 16th century than at its beginning. The lot of women, in particular, had deteriorated. Around 1500 many German women had been at work in numerous urban occupations. But a century later they had been crowded out of all but the most demeaning trades as economic pressures, reinforcing ancient prejudices, eliminated them wherever they offered competition to male craftsmen. In this light, it is not surprising that the period from the 1580s to the 1620s also witnessed a surge of persecutions for witchcraft in Germany (mainly in the southwest and Bavaria). As

elsewhere, the witch craze in the empire seems to have been a reaction to the strains of a time of troubles, the actual causes of which, fairly clear now to historians, were hidden from contemporaries.

Religion and politics: 1555–1618. Four forces contended for supremacy in the Holy Roman Empire in the aftermath of the Peace of Augsburg. Lutherans—that is to say, Lutheran estates and governments—sought to extend the rights they had won in 1555 to parts of Germany still Roman Catholic. Calvinists, having been excluded from the Augsburg settlement, strove for recognition and made major territorial gains in the 1560s and '70s. Adherents of the old faith, invigorated by the Catholic Reformation issuing from Spain and Rome, attempted to turn back the Protestant advance by making common cause with strong governments. Habsburg emperors tried to serve the Catholic cause by weakening Protestant princes wherever possible and by holding the line against Protestantism in their dynastic lands. Political conflicts were constant under these circumstances and wars frequent, the empire's institutions being powerless to neutralize or channel these competing endeavours. Maximilian II from 1564, Rudolf II from 1576, and Matthias from 1612, though ardent Catholics, were preoccupied with the intertwined problems of retaining the loyalty of their dynastic lands and securing the eastern borders against the Turks. They had, in any case, been stripped of the ability to maintain order in the empire, as the Augsburg terms had placed public security under the supervision of the empire's administrative districts, which were controlled by the estates. The period leading up to the Thirty Years' War was therefore one of more or less constant strife in nearly all parts of the empire.

The second half of the 16th century introduced two new agents of change to this scene. The Catholic Reformation, operating mainly through the Council of Trent (1545–63) and the Society of Jesus, brought about major changes in Roman Catholicism. Trent produced authoritative definitions of dogma for the first time in the long history of the church, declared tradition to be, with the Bible, a source of revelation, reaffirmed the sacraments as mediators of grace, declared the church to be a hierarchical institution headed by the pope (against Luther's formulation of the "priesthood of all believers"), and issued a large number of reform mandates to meet, at last, the age-old charges of laxness and corruption. After the 1560s the Catholic Reformation's chief energies went to the implementation of the Trent decrees. Most effective in this endeavour were the Jesuits, a militant order founded by Ignatius of Loyola in 1534, pledged to strict obedience to the pope and to acting as the church's instrument for regaining ground lost to Protestantism. Germany was a major area of Jesuit activity; the order settled in Cologne in 1544 and later in Vienna, Ingolstadt, and Prague. In close collaboration with Catholic rulers, often as their confessors, the Jesuits embodied the activist phase of Catholic reform that goes under the name of Counter-Reformation.

The Counter-Reformation

On the Protestant side, this activism was represented by the Calvinists, who made so forceful an impact on German society in these decades that some historians have called their appearance a "Second Reformation." The Palatine Electorate went Calvinist when its ruler converted; later the "Reformed" creed (as its partisans named it, denying to other Protestant denominations the claim to have truly reformed the faith) established itself, among other places, in the electorates of Brandenburg and (for a time) Saxony, the territories of Hessen-Kassel, Nassau, Durlach, and Anhalt, and the cities of Bremen, Emden, and Münster. Unlike Genevan Calvinism, the Reformed religion in Germany coexisted easily with the autocratic territorial church; German princes, for their part, saw Calvinism as a far more aggressive theological and political weapon with which to wage the struggle for Protestant supremacy in the empire. Calvinist theology, with its emphasis on action in the world and its association of success with sanctification, even with election, was well suited, in the use made of it by German state churches, to act as an aggressive creed of social discipline; it also inspired the formation of militant parties pressing for recognition of Calvinism as a legitimate religion under the *cuius regio, eius religio* rule.

With the religious situation thus more inflamed than ever and the confessional and political issues inextricably intertwined, any incident might have triggered renewed conflict, which—given the competition for power in Europe among the Habsburg dynasties, France, England, and the Netherlands—was likely to lead to a general war. A series of incidents moved events toward the brink. In 1582 the archbishop-elector of Cologne, having converted to Calvinism, challenged the Ecclesiastical Reservation of the 1555 Augsburg treaty by holding on to his title, thus threatening to throw the majority vote in the College of Electors to the Protestants. In the "Cologne War" of 1583 he was expelled by Spanish troops, and Duke Ernst of Bavaria was chosen as his successor. Throughout the 1590s the incorporation of church properties by Protestant governments was a cause of litigation before the empire's courts, as Roman Catholic authorities sought to compel the return of everything confiscated since 1555; Protestant estates, in turn, made support for the emperor's war against the Turks dependent on further concessions.

The Habsburgs, meanwhile, were hampered in their advancement of the Roman Catholic cause by the growing mental incapacity of Rudolf II; indeed, much of the direction of affairs was transferred to his brother Matthias, who eventually succeeded him in 1612. A more serious undermining of Habsburg imperial power occurred in the dynastic lands. Rigorous Catholic reform occasioned peasant uprisings in Austria and resistance by nobles in Hungary and Bohemia (where Calvinism had made inroads among the ruling classes). In Hungary a nationalist party under István Bocskay forged an alliance with Lutheran princes and obtained support from the Turks. In Bohemia, in 1609, the estates extracted from the emperor a guarantee of religious freedom, the so-called Letter of Majesty.

At about the same time the city of Donauwörth was occupied by Bavarian troops, Duke Maximilian of Bavaria having been empowered by the emperor to "protect" the Roman Catholic minority there. Seeing this "Donauwörth incident" as a straw in the wind, Lutheran and Calvinist rulers formed a Protestant Union (1608), the answer to which was the Catholic League (1609), headed by Maximilian, the most resolute Catholic prince in the empire. Each party organized an army and allied itself with foreign powers, the Protestants with France and Bohemia, the Catholics with Spain. In this way the German struggle was both militarized and internationalized.

General war nearly broke out in 1609–10 over the Jülich-Cleves succession crisis. When the Roman Catholic ruler of these counties, which formed the strategically most important block of territories on the lower Rhine, died without issue, two Protestant claimants occupied his lands, aided not only by the German Protestant Union but also by France and England; they were, however, militantly opposed by Spain and the emperor. The assassination of Henry IV of France, who had been about to launch an invasion in support of the Protestant claimants, defused the crisis in 1610.

The Jülich-Cleves succession crisis

Peace was preserved, although not for long. The Bohemian situation finally precipitated the war. Because neither Rudolf II nor Matthias left legitimate heirs, the governance of the Habsburg dynastic lands fell to Archduke Ferdinand of Styria (later Emperor Ferdinand II), a ruthless counter-reformer who reduced the religious liberties granted to Bohemians under the Letter of Majesty. In response, the Bohemian estates in May 1618 mounted a protest in Hradčany (Prague Castle), which prompted a militant faction of deputies to throw two imperial councillors from a castle window ("defenestration" being a traditional Bohemian gesture of defiance). Ferdinand now prepared military action, while the Bohemian estates elected a Calvinist, Frederick V of the Palatinate, to be their king. As the alliances fell into place on each side, the stage was set for the sequence of large-scale military actions that constituted the Thirty Years' War.

Defenestration of Prague

The Thirty Years' War and the Peace of Westphalia. The Bohemian problem was resolved swiftly. Two Roman Catholic armies, the emperor's and the League's, converged on the kingdom, routing Frederick at the White

Mountain in November 1620 and replacing the regime of the estates in Bohemia with a system of "confessional absolutism" based on rigid Catholic conformity and political authoritarianism. At the same time, the Palatinate was conquered by Spanish and Bavarian troops, and the electoral title, forfeited by Frederick, transferred to Maximilian of Bavaria in 1623. In the Palatinate, too, the Counter-Reformation sought to bring Protestantism to an end. As the war spread into Hesse and Westphalia, and as Spain resumed its attack on the Netherlands, Catholic forces seemed near triumph. This prospect, however, renewed the old fear, as in the time of Charles V, of Habsburg hegemony; an anti-Habsburg alliance was therefore forged by France (where Cardinal Richelieu took charge of affairs in 1624), England (James I was father-in-law to the deposed Frederick V), the Netherlands, and Denmark (whose Protestant king, Christian IV, had extensive territorial interests in north Germany, now threatened by Catholic armies). In 1625 Christian commenced hostilities. He was opposed by a much-enlarged imperial force under the war's most flamboyant figure, Albrecht von Wallenstein, a military entrepreneur with a gift for mobilizing troops and supplying them through ruthless plundering. Wallenstein's plan was to wreck Dutch and English commerce in the Baltic by subduing all of north Germany and by dislodging the Swedish king, Gustav II Adolf, from Prussia (taken in the course of Sweden's war against Poland). By 1628 much of this objective was realized, Christian IV having been decisively defeated in 1626 in the Battle of Lutter am Barenberge and forced to make peace in 1629. To crown these successes, Ferdinand II issued in March of that year an Edict of Restitution, by which Protestant rulers were to restore to the church more than 500 bishoprics, monasteries, abbeys, and other ecclesiastical properties secularized since 1552.

But this impressive strengthening of the sovereign's power in the empire brought his foreign and domestic enemies together once more, the latter including now not only Protestants but also Roman Catholic estates concerned about their liberties. Subsidized by the Dutch and by France and allied with Saxony, Sweden entered the conflict in 1630, winning commanding victories at Breitenfeld (1631) and Lützen (1632) but suffering defeat at Nördlingen in 1634. This phase of the war was marked by unprecedented brutality; for example, in 1631, imperial troops massacred two-thirds of the population of Magdeburg, a city of 20,000 that had withstood a long siege.

The Magdeburg massacre

A way out of the long conflict appeared in 1635 when Saxony, Brandenburg, and other Protestant estates seeking to end foreign intervention joined the emperor in the Peace of Prague, which included the revocation of the Edict of Restitution. But in this, the war's final phase, the initiative passed to France, which, seeking to forestall Spanish preponderance on the Continent, offered large subsidies to Sweden and to German princes to enable them to fight on. Combined Swedish-French campaigns commenced in 1638, and a decade later foreign armies operated as far south as Bavaria, while the French held Lorraine and Alsace, which was important to France to prevent construction of a Spanish land bridge from the Netherlands to Italy.

By then most belligerents were exhausted. Several German princes had quit the war. Since 1644, representatives of the powers had been talking about terms, although military operations continued in hopes of improving bargaining positions. In 1648, finally, treaties were signed in Münster and Osnabrück (both in Westphalia) by agents of the emperor, the German estates, Sweden, and France as well as between Spain and the Netherlands. Fighting continued for some years—France and Spain did not conclude peace until 1659—but the war was at last winding down.

The Treaties of Westphalia

The Peace of Westphalia brought territorial gains to Sweden and France, awarded an electoral seat to Bavaria, and secured for Protestant rulers the church properties they had confiscated, based on the status quo of 1624. More important, it brought Calvinists into the religious settlement and established the independence of the Netherlands from Spain and of Switzerland from the empire. Most significant of all, it guaranteed the nearly unlimited territorial sovereignty of German princes, bringing to an end the last effort (until the 19th century) to turn the empire into a centrally ruled "modern" state. In this way 1648 sealed the fragmentation of the Holy Roman Empire into hundreds of autonomous political entities, most of them small. At the same time, it brought to an end the last major conflict in continental Europe in which religion was a salient issue; indeed, the war itself had demonstrated that reason of state was a stronger determinant of policy than faith. In declaring the religious situation fixed as of 1624, the treaty eliminated the *jus reformandi* as a cause for confessional change; if a prince converted, his land no longer converted with him. Religious pluralism and—albeit grudgingly—coexistence were now the norm.

The war's social and economic cost is difficult to gauge, modern scholarship having greatly modified original claims of vast human losses and near-total economic ruin. Nonetheless, in the most embattled realms, such as Württemberg, more than 50 percent of the people died or disappeared; elsewhere, the loss was less severe. Most historians agree that an overall population decline of 15 to 20 percent (from *c.* 20 million to 16 or 17 million) occurred during the war and the ensuing epidemics. In addition, historians agree that in theatres of war rural impoverishment and displacement of people were widespread, while economic regression happened nearly everywhere. For German society overall, the war was a traumatic experience; it is rivaled in the national consciousness only by the 1939–45 war as a time of unmitigated disaster.

To gain perspective on these calamities, their wider European aspects must be considered. Wars, uprisings, and political turmoil had occurred in many countries during the first half of the 17th century. The Fronde—a series of civil wars in France between 1648 and 1653 whose goal, at least in part, was to halt the growing power of royal government—and the Civil War in England (1642 to 1651) are only the most famous of these disturbances. Turmoil had occurred also in Catalonia, Portugal, Naples, Ireland, Scotland, Sweden, and Russia. Historians have referred to these events, including the numerous local manifestations of the Thirty Years' War, as parts of a general crisis in the fabric of European society, the causes of which range from a worsening of the climate (Little Ice Age) to plagues, often spread by the armies roaming Europe almost continuously at that time. But the most destabilizing factor burdening society was the centralizing monarchy with its expanding bureaucracy, extravagant courts, swollen armies, and incessant wars, all of them supported by heavy taxation. No social group, or estate, was unaffected by the effort of monarchs to alter in their favour traditional ways of distributing power and influence. Resentment of this and of its social cost was widespread; hence the proliferation and the scale of rebellions.

TERRITORIAL STATES IN THE AGE OF ABSOLUTISM

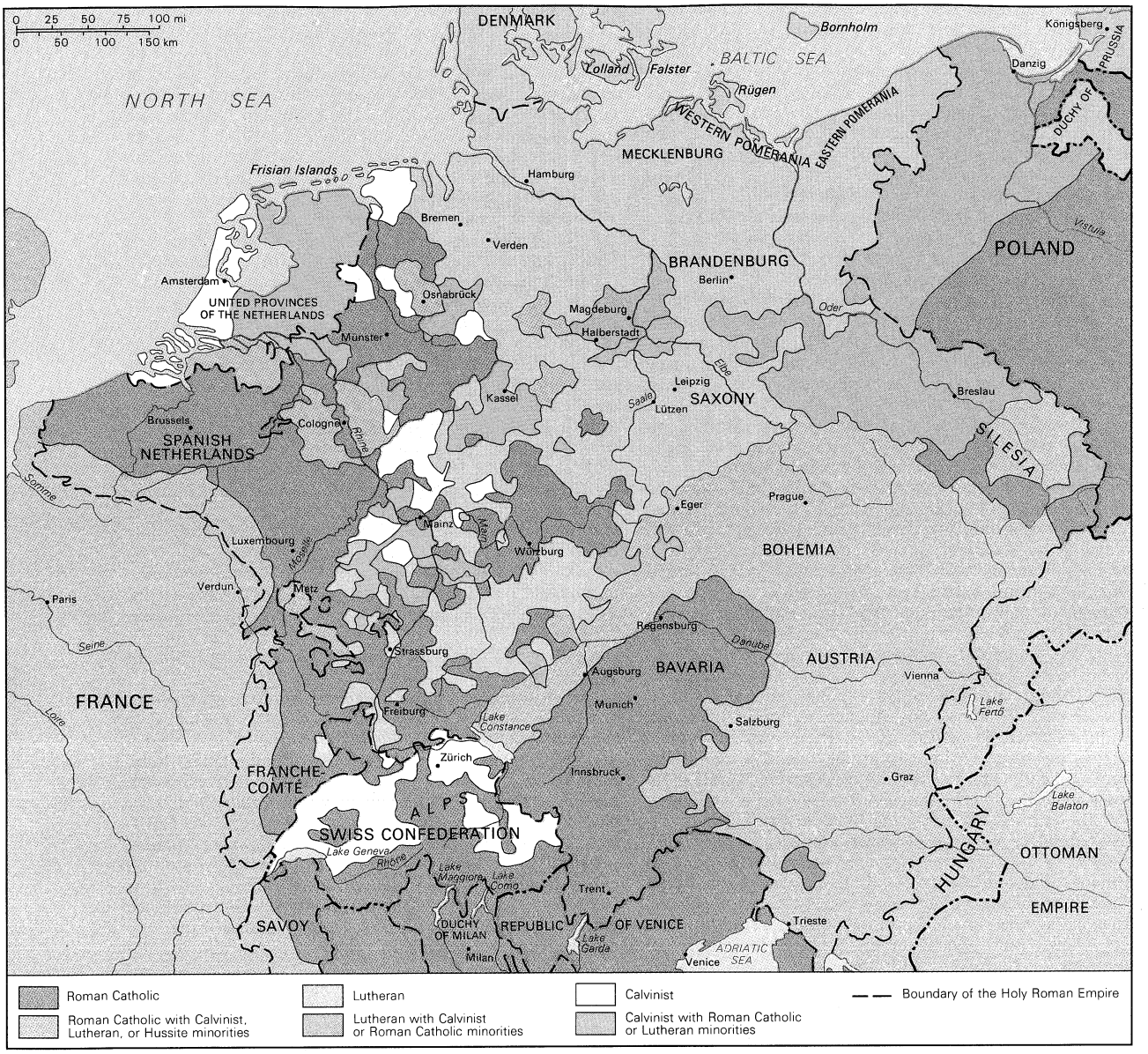
The empire after Westphalia. The empire was an awkward structure. German historians of an older, nationalistic generation deplored the fact that the empire lacked any of the attributes of a great power and lamented its victimization by more unified foreign states. Such critics always quote the 17th-century legal scholar Samuel Pufendorf, who called the empire a "monstrosity," and interpret this term as a value judgment rather than an expression indicating the inapplicability of standard categories of political classification. Recent scholars have been more appreciative of the post-1648 empire as a loose-jointed but not ineffective constitutional edifice within which could coexist 300 large and small, secular and ecclesiastical principalities, 51 imperial (*i.e.,* immediate to the empire) cities, and nearly 2,000 imperial counts and knights, each of whom possessed the same territorial sovereignty as an elector or a duke. The empire proved a working federation for the varied interests of these distinct sovereign entities, some of them large and powerful (like Saxony, whose electors were also kings of Poland from 1697 to 1763, or Brandenburg, whose prince was also king in Prussia) and some laughably tiny (such as the Abbey of Baindt in Swabia, a fully independent territory of a few hundred acres inhabited by 29 nuns and governed by a princess-abbess).



Germany in 1648.

The empire's administrative organs, especially the districts (*Kreise*), protected the small and weak from the predatory aims of the strong. Because most constituent members were vulnerable, there was no general inclination, despite disunity among the estates on matters of taxation and religious parity, to break the frame that guarded the status quo. The emperor's suzerainty over the entire realm went unchallenged, but virtually no real power adhered to his title, executive authority having been thoroughly particularized between 1555 and 1648. To prevent a resurgence of imperial power, princes formed alliances among themselves, such as the League of the Rhine (Rheinbund), tied in 1658 to France and Sweden. In the princely territories authority fell increasingly to the princes (exceptions were Württemberg and Mecklenburg), while territorial estates dwindled in political importance. In each of the empire's constituent units, estates served mainly to uphold established hierarchies and traditions, as did the empire as a whole. It was an inherently conservative system.

The consolidation of Brandenburg-Prussia and Austria. Against an overall tendency among the empire's constituent units to keep things as they were, the larger territories pursued an insistent policy of dynastic and personal aggrandizement. A number of factors favoured state building in the post-1618 era. General economic exhaustion made central direction of, and active intervention in, commerce and industry seem to be the only way out of stagnation. War taxes, raised to a steep level during the French wars of the 1670s (see below), greatly increased the financial might of rulers, who came to control an unprecedented share of society's wealth by preparing for and engaging in military conflict. Because territorial assemblies opposed this siphoning process—whose proceeds, augmented by subsidies from abroad, served mostly to create standing armies and a supporting state apparatus—rulers attempted to reduce even further the estates' role in policymaking. The nobility, growing economically dependent on princely service, adapted itself to an essentially ancillary function at court. In society at large the view gained ground that the country's welfare was safest with the ruler—a view vigorously promoted by official propaganda. Two of the empire's territories, Brandenburg-Prussia and Austria, profited above all others from these developments.



The range of confessions in Germany, 1650, as a result of the Thirty Years' War.

Frederick William, the "Great Elector"

The story of Brandenburg-Prussia has always been read as exemplifying the triumph of political skill and audacity over unfavourable conditions. Sparsely populated and deficient in resources, Brandenburg in 1648 was a patchwork of scattered territories. Its ruler, Frederick William (1640–88), later called for his achievements the "Great Elector," faced the problem of integrating and defending widely separated possessions, which included the duchy of Prussia, inherited in 1619 but remaining under Polish suzerainty and geographically separated from the electorate; the counties of Cleves, Mark, and Ravensberg in the Rhine and Westphalia regions, gained in 1614, also distant from Brandenburg and not contiguous with each other; and eastern Pomerania and various small lands and bishoprics acquired in the Treaty of 1648. Through nimble diplomatic maneuvering, such as changing sides several times between Sweden and Poland and beweeen France and the emperor, he augmented and solidified his realm and his authority within it; he won direct rule over Prussia as its duke and acquired the important episcopal territory of Magdeburg.

Frederick William's instrument in the attainment of these and subsequent prizes was the army, a permanent force of 30,000 disciplined professionals, the adequate financial support of which dictated every aspect of his government. Large revenues from taxes required a flourishing economy, the stimulation and direction of which by mercantilist principles was a main undertaking. Economic growth was further accelerated late in the Great Elector's reign by the influx of nearly 20,000 skilled Huguenot refugees following the revocation of the Edict of Nantes by Louis XIV in 1685 and by the resettlement of Dutch colonists. A territorywide system of state administration undergirded this economic and fiscal effort and resulted in the creation of a professional bureaucracy that permitted the Great Elector to govern essentially without estate participation. The land-owning nobility supported their prince in exchange for the freedom to exploit their peasants as they saw fit. In these ways Frederick William laid the foundation for what was to become an autocratically ruled state, enabled by its strong economy, tightly run administration, efficient fiscal organization, and powerful army to play a prominent role in the empire's and Europe's affairs.

"King in Prussia"

The Great Elector's efforts were rewarded in 1701 when his successor, Frederick III (1688–1713), obtained from the emperor (who needed the Brandenburg army for the impending War of the Spanish Succession) the right to style himself "King in Prussia." The title, recognized internationally upon the conclusion of the war in 1713, was of considerable importance to Brandenburg in its competition with Saxony, whose ruler had become king of Poland in 1697, for preeminence in north Germany. But it was Frederick's son, Frederick William I (1713–40; Prussian rulers renumbered themselves upon bestowal of the royal

title) who perfected the combination of statist structure, productive energy, and ethical drive that came to be identified with modern Prussia. Known as the "soldier-king," Frederick William built his standing army into a force of more than 80,000 men. It was the fourth largest army in Europe (after those of France, Russia, and Austria) in a country only thirteenth in population, superbly drilled and equipped but serving mainly defensive purposes. Only peasants and journeymen served in the ranks, many of whom were impressed abroad, while the middle classes were safe from the draft but obliged to quarter soldiers in their homes. A huge war chest obviated foreign subsidies, and reliable revenues, more than 70 percent of which went to the army, provided ample support.

To continue to draw high taxes without ruining land and people, the country's level of wealth had to be raised. Frederick William therefore pursued an aggressive mercantilist policy of stimulating agriculture and manufacturing while reducing unnecessary expenditures; even his court was stripped of many of its royal trappings. Export bans preserved raw materials, and sumptuary laws limited indulgence in luxuries. Town governments ceded authority to royal commissioners, whose powers included supervision of urban production. A work ethos suffused society from the top; the king's ascetic Calvinism, which dictated to him a life of hard work and personal engagement, was spread to his Lutheran subjects by a Pietist clergy who instilled in their flocks habits of intense labour, frugal living, and dutiful subservience to the state.

Organizationally, Frederick William completed the centralizing process begun by the Great Elector, its capstone being the General Directory, set up in 1723. Tied to regional and local organs by a network of commissioners, this supreme body of state policy and administration directed industry, trade, finance, internal affairs, and military matters in all the state's territories. Upper-level bureaucrats came entirely from the nobility, as did the army's officer corps; in this way nobles were bound more closely than ever to the state. Ruling, not merely reigning, over the entire edifice was the king-elector in his "cabinet," a small circle of close advisers and trusted secretaries. So successful were these measures in lifting the state to influence and prestige that by 1740 Prussia counted as a full-fledged member of Europe's concert of great powers.

In Austria, the ruling Habsburg house's lasting conflict with France and the Ottoman Empire dominated all questions of statecraft. With their powers as emperors greatly diminished, Leopold I (1658–1705), his son Joseph I (1705–11), and Joseph's brother Charles VI (1711–40) bent all their efforts to the consolidation of their dynastic and crown lands in central and eastern Europe. Although they failed to achieve Prussian-style streamlining, they raised Austria to the rank of a major state. The Habsburgs' conglomeration of territories included the Austrian duchies (Austria, Styria, Carinthia, Krain, and the county of Tirol), the Bohemian provinces (kingdom of Bohemia, Moravia, Silesia), the kingdom of Hungary, and—after 1714, following the War of the Spanish Succession—the southern Netherlands (Brabant, Luxembourg, Flanders) and the duchy of Milan.

These disparate lands were held together only by the Habsburg monarchy, but the monarchs were distracted from the task of integrating them. They were preoccupied by imperial concerns and by dynastic complications, notably the succession question. Until the reforms of Maria Theresa's reign (1740–80) Austrian administration never became effective. Finances were especially muddled, because tax administration remained with the estates of the various territories, along with control over other sources of revenue. The army of 100,000 men, though the third largest in Europe, was barely adequate for the defense of so large and scattered a realm. A supreme war council and a central financial chamber overlapped with special commissions created by the emperor's privy council, which also handled military and fiscal affairs. Nonetheless, the realm held together.

The prospect of a succession without a male heir, however, presented the severest test to the realm's cohesion. It became the chief enterprise of Charles VI to persuade the estates of his territories to accept an order of succession, known as the "Pragmatic Sanction," by which the Habsburg lands were declared indivisible and Charles's oldest daughter, Maria Theresa, was to inherit them. The other European powers assented, because splitting the Habsburg complex would have thrown the European balance into disarray and played into the hands of France; the Sanction was proclaimed a basic law in 1713. By then Austria had met successfully a series of Turkish incursions from the east and French invasions in the west.

The Pragmatic Sanction

The age of Louis XIV. For the empire as a whole, the half century following the Peace of Westphalia was almost entirely shaped by the dominant political figure of the time, King Louis XIV of France. The response of the empire and its members to the aggressive undertakings of this monarch, whose aim from his assumption of power in 1661 to his death in 1715 it was to make France the mightiest state in Europe, was largely reactive (for a different interpretation, see the article FRANCE: *The age of Louis XIV: Foreign affairs*). Only in its struggles against Louis's ally in the east, the Ottoman Turks, did the empire show some initiative. After a Polish relief army had helped imperial, Bavarian, and Saxon troops to lift a three months' Turkish siege of Vienna in 1683 in the Battle of Kahlenberg, imperial armies took the offensive, winning battles at Ofen (1686), Mohács (1687), and, most notably, Zenta (1697). In the Treaty of Carlowitz (1699) Austria gained parts of Hungary, Transylvania, Slavonia, and Croatia, all of them formerly occupied by the Turks. The eastern wars resumed in the early and mid-18th century, but the Turks were never again a threat to Europe, since Russia became the chief bulwark against Ottoman expansionism.

Matters were different on the empire's western and southern fronts. The overriding political question in Europe in the second half of the 17th century was the future of Spain and its vast holdings in the southern Netherlands, Italy, and America, because it was expected that the Spanish Habsburg line would die out with the feeble Charles II. Contenders for the Spanish inheritance were the Habsburg emperor, Leopold I, husband of a younger Spanish princess, and Louis XIV. The French monarch, son of the eldest daughter of Philip III, had further strengthened his claim to the Spanish throne in 1659 when, in accordance with the Peace of the Pyrenees, which had ended the long conflict between Spain and France, he had married the Spanish infanta.

While waiting for the Spanish throne to become vacant, Louis pursued an aggressive expansionist policy. He pushed his forces toward Germany to make the Rhine River France's new eastern border. In 1667 he occupied Flanders and in 1670 Lorraine; in 1672 he attacked Cleves and invaded the United Provinces of the Netherlands, his main antagonist in the wars that followed. In 1679 he began to penetrate Alsace, occupying the imperial city of Strassburg (Strasbourg) in 1681. Lacking the military power to bring the whole empire to its knees, Louis resorted to the lure of money; at one time or another almost every German state was in his pocket, either serving as ally or remaining neutral. Though not incapable of acting on national impulses, German princes—the Great Elector being a case in point—always served territorial interests first. This prevented the emperor, himself at times allied with Louis, from forging a solid front against France.

Leadership of the anti-France coalition passed to the Netherlands. William of Orange, as stadtholder of Holland and captain general of the United Provinces, emerged as the most determined opponent of French aggression. Upon becoming king of England in 1689, he changed the direction of English politics, which had been pro-French under the last Stuart king. The threat of a French universal monarchy arose dramatically as the death of the last Habsburg in Spain approached (Charles II died in 1700) and Louis's plans for a French claim on the entire Spanish inheritance swung into place. A Grand Alliance now formed against him (it was formally concluded in 1701), consisting of the empire (except Bavaria and the electorate of Cologne), the Netherlands, England, Sweden, Brandenburg-Prussia, and Savoy. Its aim was to restore

War of the Spanish Succession

the European balance to the status of 1648 and 1659 by ejecting Louis from his conquests and by splitting the Spanish empire.

The first phase of the ensuing struggle, known in Germany as the Palatine War, was fought in Germany. It led to savage destruction in the Palatinate and in Swabia but to no decisive victories on either side. A temporary peace (Rijswijk, 1697) forced Louis to make concessions and perhaps to realize the limits of his strength; his revocation of the Edict of Nantes in 1685 had set Protestants everywhere against him. The long-prepared War of the Spanish Succession broke out in 1701. The empire played a minor role in this conflict, Louis having failed in his diplomatic effort to enlarge the simultaneous Northern War of Sweden against Denmark, Russia, and Saxony into an attack on Brandenburg-Prussia. Austria, however, proved a major antagonist in this last attempt by the *grand monarque* to gain the upper hand. While Austrian and English troops operated in Italy, French and Bavarian forces fought in southern Germany. Neither side won clear victories, but the alliance gradually gained the advantage, until the death of emperor Joseph I in 1711 placed his brother Charles, who had been proclaimed Spanish king, on the imperial throne as Charles VI (1711–40). This raised the spectre of a Habsburg reunion of the Holy Roman and Spanish empires, which was no more agreeable to European powers than the prospect of French overlordship. Thus the alliance severed.

Peace negotiations began in 1712, resulting in a number of treaties, signed at Utrecht and Rastatt in 1713–14. The Spanish empire was partitioned, with the Spanish Netherlands, Milan, Naples, and Sicily going to Austria and Spain itself coming under the rule of Philip V of Bourbon, a grandson of Louis XIV. The alliance's original aim, to prevent French hegemony, had been achieved, though in the follow-up War of the Polish Succession (1733–35) France acquired control of Lorraine. Austria profited substantially in territorial terms, and a few other German rulers profited as well, albeit less so. As for the empire itself, it had gained no real benefits from more than half a century of intermittent warfare.

German society, however, was deeply affected. Economic stagnation and slow demographic recovery after the Thirty Years' War made Germany dependent on governmental intervention as a means of stimulating recovery. This left the country exposed to foreign influences, which reached land and people by way of the many princely courts and the elites clustered there.

German cultural life took its cues from abroad: Baroque art—the preeminent expression of monarchical power and of Roman Catholic resurgence after the Reformation—came from Spain and Italy, opera from Italy, and polite language and manners from France. The style of this period took French patterns as its model, from elaborately coded court ceremonials to dress, social conventions, food, and conversation. French absolutism not only became the political model, however scaled down, for the governance of all states in the empire, but every German prince and princeling imitated the lavish display with which Louis XIV created his aura of majesty and outshone his rivals. This started up a lively domestic market in luxuries, not to mention splendid works of architecture and decoration. But the cost of these luxuries was prohibitive (in 1719 the Palatine court consumed 50 percent of the territory's revenues) and represented an enormous burden on the people, especially when added to the cost of large armies and proliferating bureaucracies. Not only did this conspicuous consumption widen the social division between the court-oriented elite and the bulk of the urban and rural population but the culture's foreign provenance of the goods also inhibited creative impulses at home.

In the second half of the 17th century German energies were to a large extent still focused on religion. The confessional pluralism legitimized by the settlement of 1648 encouraged emphasis on theological distinctions, exacerbating the move toward religious orthodoxy under way in each denomination since the 16th century. The one genuinely German product of this religious preoccupation was Pietism. It was a religious movement within Lutheranism that opposed rigid dogmatism and promoted instead a subjective, mystical devoutness and an emphasis on a pious life guided by love of one's neighbour as well as of God. Influenced by English Puritanism, Pietism was shaped in its theology by Philipp Jakob Spener (1635–1705) and in its organization by his disciple August Hermann Francke (1663–1727), who established a centre for its promulgation in the Brandenburg city of Halle. There he founded schools, orphanages, medical facilities, and a printing house for publishing cheap Bibles and devotional works, which made of Pietism a widely influential program of evangelical activism. The intensely emotional and mystical flavour of Pietist poetry is preserved in the cantata texts set to music by Johann Sebastian Bach, in whose deeply spiritual church music Protestant chorale singing, another indigenous German product, reached its apogee.

The contest between Prussia and Austria. In 1740 the death of the Habsburg emperor Charles VI without male heir unleashed the most embittered conflict in Germany since the wars of Louis XIV. The question of who would ascend the Austrian throne had occupied statesmen for decades. Rival claimants disputed the right—by the terms of the Pragmatic Sanction (1713)—of Charles's daughter Maria Theresa to succeed; France supported them, its aim being, as before, the fragmentation of the Austrian state. But it was the new Prussian king, Frederick II (1740–86), who began the conflict. To understand what follows, the modern reader should remember that few observers, even in the enlightened 18th century, disputed a ruler's right to do what he wished with his state. Dynastic aggrandizement, territorial expansion, prestige, honour, power, and princely glory were legitimate grounds for war and sound reasons for demanding the sacrifices necessary to wage it. The only position from which to oppose this arrogation was the Christian ethic, but to do so had proved futile when last tried by Erasmus and Sir Thomas More in the 16th century. No checks—philosophical, moral, or political—therefore restrained kings from indulging their taste for conquests.

Soon after taking control of affairs, Frederick reversed his father's cautious policy of building and hoarding, rather than consuming, Brandenburg-Prussia's military potential. He attacked Silesia, a province in the kingdom of Bohemia and thus part of the Habsburg monarchy, which Prussia had long desired for its populousness, mineral resources, and advanced economy. In exchange for an Austrian cession of Silesia he offered to accept the Pragmatic Sanction (formally recognized by his predecessor in the Treaty of Berlin in 1728) and support the candidacy of Maria Theresa's husband, Francis Stephan, as emperor. But the resolute woman who now headed the Austrian Habsburgs (1740–80) decided to defend the integrity of her realm, and war began in 1740. Austria was helped only by a Hungarian army, though financial support came from Austria's ally, England. Prussia was joined by Bavaria and Saxony in the empire as well as by France and Spain. The Prussian armies, though greatly outnumbered against forces that included troops from Russia and Sweden, revealed themselves as by far the best as well as the best led. In the Treaty of Hubertusburg of 1763 Prussia kept Silesia but could not hold on to Saxony, which Frederick had invaded in 1756 and which he coveted as a way of rounding out his state.

War of the Austrian Succession

In a sense, the War of the Austrian Succession (1740–48) was another of the many internal struggles over the constitutional balance in the empire in which territorial states opposed imperial authority. But it was also part of an international struggle, with France and England fighting out their rivalry in western and southern Europe, America, and India. Thus it blended into the worldwide Seven Years' War (1756–63), the latter being preceded by a "Diplomatic Revolution" in which England switched its support to Prussia and France allied itself with its traditional foe, Austria (a part of this agreement was the marriage, in 1770, of the Austrian princess Marie Antoinette to the future Louis XVI). The real significance of the Seven Years' War lay in the Peace of Paris of 1763, which concluded for a time the maritime and colonial conflict between France and England.

Austria having survived the war and Prussia having increased in size and immeasurably in prestige, these two powers now dominated German affairs in a condition of tension usually called "the German dualism," meaning that each had become so powerful that only the other could keep it in some sort of check. The monarchs of both realms carried out important internal reforms. Guided by her interior minister, Count Friederich Wilhelm Haugwitz, Maria Theresa streamlined the Austrian administrative structure on the Prussian model, thus drawing together, to the extent possible, the multiethnic and polyglot regions of the far-flung Habsburg empire. The remaining powers of the estates were curtailed everywhere and centralization institutionalized in absolutist fashion but without attaining the full integration of the Prussian system. Maria Theresa's son, Joseph II (1765–90), completed this program of modernization.

In Prussia, Frederick II further tightened his control of all aspects of public life in his far-flung kingdom. But in accordance with his personal commitment to rational tolerance and free-thinking skepticism he also undertook extensive legal reforms; he virtually abolished judicial torture, lifted some of the tax burden from the poorest of his subjects, established religious tolerance as a policy of his state, and encouraged scientific and scholarly activity in the Prussian Academy. Like his father, he was a vigorous promoter of economic development and colonization. His taste for French Enlightenment thought and his own prolific creativity in letters and music lent to his reign the flavour of an era shaped by a philosopher-king, albeit one with the instincts of a ruthless power politician. His successes in war and peace earned him a place as national hero as well as the title "the Great." (Ge.St.)

Germany from *c.* 1760 to 1871

GERMANY TO 1815

Germany in the middle of the 18th century was a country that had been drifting in the backwaters of European politics for more than a hundred years. The decisive roles in the affairs of the Continent were played by those great powers—such as France, England, and Spain—whose economic resources and commercial connections provided a solid foundation for their military might. The states of central Europe, on the other hand, floundered in a morass of provincialism and particularism. All the forces that had contributed to the rise of powerful national monarchies west of the Rhine were lacking in the east. In the Holy Roman Empire the central government was losing rather than gaining strength, the princes were enlarging their authority at the expense of the crown, and business initiative was being discouraged by the lack of political unity and by the remoteness of the major trade routes.

The Holy Roman Empire

Civic power became increasingly concentrated in the hands of local governments controlled by aristocratic overlords, ecclesiastical dignitaries, or municipal oligarchs. The history of Germany between the Thirty Years' War and the French Revolution is largely the sum total of the histories of dozens upon dozens of small political units, each enjoying virtually full rights of sovereignty. The rulers of these gingerbread principalities, copying the example of the royal court of France or Austria, built costly imitations of the palaces of Versailles and Schönbrunn, which today are the delight of tourists but which were once the curse of an impoverished peasantry. The tradition of princely authority, an instrument of national greatness in western Europe, encouraged national divisiveness in central Europe. The petty rulers of Germany legislated at will, levied taxes, concluded alliances, and waged wars against each other and against the emperor. Policies pursued in Munich, Stuttgart, Dresden, or Darmstadt reflected policies originating in Paris, Vienna, London, or Madrid but without seeking a goal greater than the promotion of particularistic interests.

Political institutions designed theoretically to express the will of the nation continued to function, yet they had become an empty shell. The Holy Roman emperor was still elected in accordance with a time-honoured ritual that proclaimed him to be the successor of Caesar and Augustus. The splendid coronation ceremony in Frankfurt am Main, however, could not disguise the fact that the office conferred on its holder little more than prestige. Since all the emperors except Charles VII were Habsburgs by birth or marriage, they enjoyed an authority that had to be respected. But that authority rested not on the prerogative of the imperial crown but on the possession of hereditary lands stretching from Antwerp in the west to Debrecen in the east. The sovereigns of the Holy Roman Empire, in other words, were able to play an important role in German affairs by virtue of their non-German resources. And, since Germany was not the main source of their strength, Germany was not the main object of their concern. The emperors tended to regard the dignity bestowed upon them as a means of furthering the interests of their dynastic holdings. The Imperial Diet meeting in Regensburg had also become an instrument for the promotion of particularistic advantage rather than national welfare. It continued in theory to express the will of the estates of the realm meeting in solemn deliberation. In fact it had degenerated into a debating society without authority or influence. The princes had ceased to attend the sessions, so that only diplomatic representatives were left to discuss questions for which they were powerless to provide answers. The other central institutions of the empire, such as the imperial cameral tribunal in Wetzlar, languished in indolence. Constitutionally and politically, Germany in about 1760 resembled Poland in that a once vigorous and proud state had become weakened by internal conflict to the point that it invited the intervention of its more powerful neighbours.

What saved Germany from the fate of Poland was the ability of one of the member states to defend the empire against aggression. For 200 years Austria acted as the bulwark of central Europe against French expansion. Its possessions, forming a chain of protective bases extending between the North Sea and the Danube, had time and again borne the brunt of attacks by Bourbon armies. The frontiers of France kept moving closer to the Rhine, but the Holy Roman Empire was at least spared the tragedy of partition that befell the Polish state. It was partly in recognition of the vital role that the Habsburgs played in the defense of Germany that the electors chose them as emperors with such regularity. The Austrian monarchy, moreover, endowed with resources comparable to those of the western nations, was able to pursue a policy of political rationalization with greater success than the principalities. The rulers in Vienna succeeded in improving the administration, strengthening the economy, and centralizing the government. Until the middle of the 18th century, Austria remained the only great power east of the Rhine.

The traditional role of the Austrian Habsburgs

Further rise of Prussia and the Hohenzollerns. The emergence of the Hohenzollerns as rivals of the Habsburgs and the beginning of the Austro-Prussian dualism created the possibility of reversing the process of civic decentralization that had prevailed in central Europe since the late Middle Ages. The interests of the territorial princes of the Holy Roman Empire inclined them toward a policy of particularism, while the government of Austria, with its Flemish, Italian, Slavic, and Magyar territories, could not perforce become the instrument of German unification. Prussia, on the other hand, was militarily strong enough and ethnically homogeneous enough to make national consolidation the main object of statecraft. But, though the creation of a united Germany became its mission, this mission was not from the outset one that it accepted willingly or even consciously. The intention of Frederick the Great and of his successors Frederick William II and Frederick William III was to pursue dynastic rather than national objectives. Like the lesser princes of central Europe, all they sought was to maintain and enlarge their authority against the claim of imperial supremacy. Far from wanting to end the disunity of Germany, they hoped to prolong and exploit it. The patriotic Prussophile historians, who a hundred years later argued that what Bismarck had achieved was the consummation of what Frederick had sought, were letting the present distort their understanding of the past. In fact, the greatest of the Hohenzollerns had remained as indifferent to the glaring

political weaknesses of his nation as to its great cultural achievements. His attitude toward the constitutional system of the Holy Roman Empire was similar to that of the self-seeking princelings who were his neighbours and from whom he was distinguishable only by talent and power. He may have scorned their sybaritic way of life, but politically he wanted what they wanted—namely, the freedom to seek the advantage of his dynasty without regard for the interests of Germany as a whole.

His preoccupation with the welfare of his state rather than with that of his nation is apparent in the strategy by which he tried to check Habsburg ambitions after the Seven Years' War (1756–63). During the first half of his reign he had relied primarily on military force to restrict and undermine imperial authority. In the second half he preferred to employ the weapons of diplomacy to achieve the same end. In 1777 the ruling dynasty of Bavaria came to an end with the death of Maximilian Joseph. The elector of the Palatinate, Charles Theodore, now became ruler over the territories of both branches of the house of Wittelsbach. Without legitimate offspring to whom to leave his state and without affection for his newly acquired eastern possessions, he agreed to a plan proposed by Emperor Joseph II according to which part of the Bavarian lands would be ceded to Austria. But any increase in the strength of the Habsburgs was unacceptable to Frederick the Great. With the tacit approval of most of the princes of the empire, he declared war against Austria in 1778, hoping that other states within and outside central Europe would join him. In this expectation he was disappointed. Yet Joseph also became discouraged by the difficulties encountered in what he had believed would be an easy success. The War of the Bavarian Succession dragged on from the summer of 1778 to the spring of 1779, with neither side enhancing its reputation for prowess on the field of battle. There was much marching back and forth, while hungry soldiers scrounged for food in what came to be called the "Potato War." The upshot was the Treaty of Teschen (May 1779), by which the Austrian government abandoned all claims to Bavarian territory except for a small strip along the Inn River. The conflict had brought Frederick no significant military victories, but he had succeeded in frustrating Habsburg ambition.

The War of the Bavarian Succession

Joseph II, however, was a stubborn adversary. In 1785 he once again advanced a plan for the acquisition of Wittelsbach lands, this time on an even more ambitious scale. He suggested to Charles Theodore nothing less than an outright exchange of the Austrian Netherlands for all of Bavaria. The emperor, in other words, proposed to surrender his distant possessions on the North Sea, which were difficult to defend, for a territory that was contiguous and a population that was assimilable. The scheme went far beyond that which Prussia had defeated seven years before, and Frederick opposed it with equal determination. He hoped to enlist the diplomatic aid of France and Russia against what he regarded as an attempt to upset the balance of power in central Europe. But, more than that, he succeeded in forming an Association of Princes, which 17 of the more important rulers in Germany joined. The members pledged themselves to maintain the fundamental law of the empire and to defend the possessions of the governments included within its boundaries. The growing opposition to the absorption of Bavaria by Austria persuaded Joseph that the risks inherent in his plan outweighed its advantages. The proposed exchange of territories was dropped, and Frederick could celebrate yet another triumph of his statecraft, the last of an illustrious career. But the association of princes that he founded did not survive its author. Its sole purpose had been the protection of princely prerogative against imperial authority. Once the danger had passed, it lost the only justification for its existence. Those nationalists who later maintained that it foreshadowed the creation of the German Empire misunderstood its origins and objectives. It was never more than a weapon in the struggle for the preservation of a decentralized form of government in central Europe.

The subordination by the Hohenzollerns of national to dynastic interests was even more apparent in the role they played during the partitions of Poland. Frederick the Great was the chief architect of the first partition, that of 1772, by which the ill-starred kingdom lost about a fifth of its inhabitants and a fourth of its territory to Prussia, Russia, and Austria. His successor, Frederick William II, helped to complete the destruction of the Polish state by the partitions of 1793 (between Prussia and Russia) and 1795 (between Prussia, Russia, and Austria). The result was bound to be an enhancement of Prussia's role in Europe but also a diminution of its role in Germany. The Hohenzollerns willingly embarked on a course that would in time have transformed their kingdom into a binational state comparable to the Habsburg empire. The German population in the old provinces would have been counterbalanced by the Slavic population in the new; the Protestant faith of the Prussians would have had to share its influence with the Roman Catholicism of the Poles; the capital city of Berlin would have found a competitor in the capital city of Warsaw. In short, the centre of gravity of the state would have shifted eastward, away from the problems and interests of the Holy Roman Empire. Yet the rulers of Prussia did not shrink from a policy that was likely to have such far-reaching consequences. They never contemplated sacrificing the advantage that their state would gain from an enlargement of its resources in order to assume the role of unifiers of their nation. Such a political attitude would have been an anachronism during the age of princely absolutism in central Europe. It was not design but accident that before long led to the abandonment by Prussia of most of its Polish possessions and that thereby allowed it to continue to play a vital part in the affairs of Germany.

Contingent results of the Polish partition

The cultural scene. Whereas in England the great literary epoch of Queen Elizabeth I had coincided with commercial and naval expansion, and in France the golden age of classicism had added lustre to the military glory of Louis XIV, German arts and letters flourished amid tiny principalities and somnolent towns that could only envy the powerful national monarchies west of the Rhine. Moreover, whereas in France and England, where public opinion could exert a significant influence on government, the debate over issues of state and society was conducted with a vigour that reflected its importance, in Germany the debate was bound to remain purely theoretical. No Voltaires, Rousseaus, or Burkes were likely to emerge out of such an environment. The thinkers of central Europe tended to emphasize introspection and spirituality. Culture became an escape from the narrow world of princely absolutism. Intellectual energies that could not reform the community fought to emancipate the individual through self-purification and self-perfection.

This was the background of German idealism, a philosophical movement seeking to liberate ethics and aesthetics from the shackles of empirical knowledge. Armed with the weapons of Kantian thought, it attempted to prove that there was a realm of experience lying beyond the categories of scientific investigation: the realm of the good, the true, and the beautiful. There were realities of the spirit and the mind, in other words, that were inaccessible to the practicality of the British empiricists or the intellectualism of the French rationalists. The disciples of idealism hoped to transcend the barriers created by nation, class, and religion. They spoke in the name of humanity as a whole, which manifested its underlying harmony through the infinite variety of its political, social, and theological categories. Gotthold Ephraim Lessing pleaded for religious toleration on the basis of a common system of ethical values to which all men of goodwill could subscribe. Johann Gottfried Herder preached that the unique character and meaning of each culture contributed to the richness of common humanity that defied state boundaries. Johann Joachim Winckelmann deified the classical ideal of beauty that he found in Greek art as an eternal standard, immune to the vicissitudes of time and history. These were views that offered an escape from the narrowness of everyday life. Men who found no scope for their talents in the petty world of princely authority could turn to the liberating spirituality of the idealist philosophy. Thought in central Europe gradually acquired a metaphysical coloration that distinguished it from the more robust pragmatism of phi-

German idealism

losophy in the west. It was during the second half of the 18th century that the Germans began to consider their country "the land of thinkers and poets."

The literary revival of the age displayed the same quality of introspective idealism as the philosophical movement. Johann Wolfgang von Goethe, the greatest genius of German letters, willingly accepted the existing system of civic and social values. He regarded the disunity of his nation as an expression of its historic character and defended the authority of the petty princes as an instrument of good government. He urged his countrymen to seek greatness not in collective action but in individual perfectibility. After a period of youthful rebellion against traditional canons of literary propriety, he turned to a classicism in which a serene acceptance of life harmonized with his own sympathy for the established order. Friedrich Schiller, a man of more turbulent temperament, felt a sense of resentment against political injustice and weakness. In his plays and poems there are occasional outbursts of indignation and appeals for reform. Yet there is also a pessimistic mood of resignation induced by the burden of civic ineffectualness that history had imposed on his people. Ultimately, he too sought refuge from the world in the poet's private vision.

The Sturm und Drang movement

The Sturm und Drang ("Storm and Stress") was a movement of literary innovation through which a group of young writers in the last decades of the 18th century sought to throw off the yoke of accepted standards of composition, but it remained confined to problems of prosody and taste, refusing to grapple with political or social issues.

The cultural achievements could not alter the harsh realities of national fragmentation and princely autocracy. They supported, however, the ideals of rational reform and social progress that the Enlightenment had introduced throughout the Continent. In Germany as elsewhere the 18th century became the age when the monarchical principle advanced the loftiest justification of its claim to power. The authority of the prince, so the argument went, was to be exercised not for his private advantage or gratification but for the greatness of his state and the welfare of his people. His power had to be unrestricted so that his benevolence might be unlimited. Absolute government was the only effective instrument for achieving the general good. Impressed by the scientific discoveries and material advances that they saw about them, men began to believe that the prejudices and injustices that had plagued society would gradually disappear before the steady march of reason.

Enlightened reform and benevolent despotism. The main source of enlightened reform was to be the crown, but many well-intentioned people of means and education also began to apply a new standard of conduct in their dealings with their fellow man. This change in attitude was apparent in the decline of religious resentments and discriminations. Never before had the relationship between Roman Catholics and Protestants among the well-to-do classes of central Europe been as free of rancor as on the eve of the French Revolution. It was at this time also that the Jews first began to emerge from the isolation to which a deep-seated intolerance had consigned them. The idea of assimilation held out to them the prospect of escape from the ghetto on the condition that they identify themselves in thought, speech, and attitude with the Christian society in which they lived. That prospect was to attract the Jewish minority in Germany more and more during the next 150 years. Religious toleration, however, was not the only article of faith of the Enlightenment. Its vision of a happier future included the reformation of education, the abolition of poverty, the alleviation of sickness, and the elimination of injustice. Men of goodwill established schools, founded orphanages, built hospitals, improved farming methods, modernized industrial techniques, and tried to raise the standard of living of the masses. While the hopes of the enlightened reformers of the 18th century far outstripped their accomplishments, the practical results of their efforts should not be underestimated.

According to the doctrines of benevolent despotism, however, the chief instrumentality for the improvement of society was not private philanthropy but government action. The state had the primary responsibility for preparing the way for that golden age which, in the opinion of many intellectuals, awaited humankind. The extent to which official policy conformed to rationalist theory depended, in central Europe as elsewhere, on the personality and ability of the ruler. Both of the leading powers of the Holy Roman Empire followed the teachings of benevolent despotism but with substantially different results. The emperor Joseph II, a well-meaning though doctrinaire reformer, attempted to initiate a revolution from above against the opposition of powerful forces that continued to cling to tradition. In the course of a single decade he tried to centralize the government of his far-flung domains, reduce the influence of the church, introduce religious toleration, and ease the burden of serfdom. His uncompromising program of innovation, however, alienated the landed aristocracy, whose support was essential for the effective operation of the government. The emperor encountered mounting unrest that did not end until his death in 1790 and the subsequent abandonment of most of the reforms that he had promulgated. Frederick the Great was more successful as an enlightened autocrat, but only because he was more cautious. His reorganization of the government was not as drastic, his belief in religious toleration remained less profound, and his assistance to the peasants did not go beyond a prohibition against the absorption of their holdings by the nobility. He invited settlers to cultivate reclaimed lands, and he encouraged entrepreneurs to increase the industrial capacity of Prussia. Among his most important accomplishments, although it was not completed until after his death, was the Prussian Legal Code, which defined the principles and practices of an absolute government and a corporative society. Yet Frederick was also convinced that the Prussian landed noblemen, the Junkers, were the backbone of the state, and he continued accordingly to uphold the alliance between crown and aristocracy on which his kingdom had been built.

Attempts at social and political reform

The achievements of benevolent despotism among the minor states of the Holy Roman Empire varied considerably. Some princes employed their inherited authority in a serious effort to improve the lot of their subjects. Charles Frederick of Baden, for example, devoted himself to the improvement of education in his margravate, and he even abolished serfdom, though without eliminating manorial obligations. Charles Augustus of Saxe-Weimar-Eisenach was a hardworking administrator of his small Thuringian principality, whose capital he transformed into the cultural centre of Germany. Charles Eugene of Württemberg, on the other hand, led a life of profligacy and licentiousness in defiance of protests by the estates of the duchy. Frederick II of Hesse-Kassel was another princely prodigal whose love of pleasure impoverished his subjects and forced his soldiers into mercenary service for England. The record of enlightened autocracy in central Europe was as uneven as in western Europe. Yet the ideas of the Enlightenment even at their best were unable to transform the basis of political life in the Holy Roman Empire. They could palliate, reform, and improve, but they could not alter a system of particularistic sovereignty and absolutistic authority resting on a hierarchical structure of society. They could not become an instrument of national consolidation or representative government. Only some great creative disruption of existing civic institutions could break through the crust of habit and tradition sanctified by history. Germany lacked the vital energies required for a process of political reconstruction. The galvanizing forces of rejuvenation and regeneration were to come from the outside.

The French Revolutionary and Napoleonic era. The French Revolution, transforming the Bourbon kingdom into a constitutional state, aroused intense excitement east of the Rhine. Most German intellectuals were at first in sympathy with the new order in France, hoping that the defeat of royal absolutism in western Europe would lead to its decline in central Europe as well. The princes, on the other hand, were from the outset fearful of the Revolution, which they regarded as a serious danger, for the example of unpunished insubordination by the French might encourage demands for reform among the Germans. The result was a growing hostility between the government in Paris and the rulers of the Holy Roman

The War of the First Coalition

Empire, which led in the spring of 1792 to the outbreak of war (the War of the First Coalition, 1792–97). The immediate occasion of the conflict was a quarrel over the rights of German princes with holdings in France and over the propagandistic activities of French émigrés in Germany. But the underlying cause was the clash of two incompatible principles of authority divided by profound differences regarding the nature of political and social justice. The course of hostilities soon revealed that the civic ideals and military tactics of the French Revolution were more than a match for the decrepit Holy Roman Empire. After 1793 the left bank of the Rhine remained under the control of France, and for the next 20 years its inhabitants were governed from Paris. Yet there is no evidence that they were dissatisfied with French rule or at least that they strongly opposed it. Devoid of a sense of nationalism and accustomed to submission to authority, they accepted their new status with the same equanimity with which they regarded a succession to the throne or a change in the dynasty. The Prussians, moreover, discouraged by defeats in the west and eager for Polish spoils in the east, concluded a separate peace at Basel in 1795 by which they in effect recognized the French acquisition of the Rhineland. The Austrians held out two years longer, but the brilliant successes of the young Napoleon Bonaparte forced them to accept the loss of the left bank in the Treaty of Campo Formio (Oct. 17, 1797).

End of the Holy Roman Empire. The peace proved short-lived, however, for at the end of 1798 a new coalition directed against France was formed (the War of the Second Coalition, 1798–1802). This time Prussia remained neutral. Frederick William III, a conscientious and modest but ineffectual ruler, was distinguishable from his father by private morality rather than political skill. The government in Berlin drifted back and forth, dabbling in minor economic and administrative reforms without introducing a significant improvement in the structure of the state. A decade of neutrality was frittered away while the army commanders rested on the laurels of Frederick the Great. Austria, on the other hand, played the same leading role in the War of the Second Coalition as in the War of the First Coalition, with the same unfortunate result. The French victories at Marengo (June 14, 1800) and Hohenlinden (Dec. 3, 1800) forced Emperor Francis II to agree to the Treaty of Lunéville (Feb. 9, 1801), which confirmed the cession of the Rhineland. More than that, those rulers who lost their possessions on the left bank under the terms of the peace were to receive compensation elsewhere in the empire. In order to carry out this redistribution of territory, the Imperial Diet entrusted a committee of princes, the Reichsdeputation, with the task of drawing a new map of central Europe. The major influence over its deliberations, however, was exercised by France. Napoleon had resolved to utilize the settlement of territorial claims to achieve a fundamental alteration in the structure of the Holy Roman Empire. The result was that the Final Recess (Hauptschluss) of the Reichsdeputation of February 1803 marked the end of the old order in Germany. In their attempt to establish a chain of satellite states east of the Rhine, the French diplomats brought about the elimination of the smallest and least viable of the political components of central Europe. They thereby also furthered the process of national consolidation, since the fragmentation of civic authority in the empire had been a mainstay of particularism. That it was not Napoleon's intention to encourage unity among his neighbours goes without saying. Yet he unwittingly prepared the way for a program of centralization in Germany that helped to frustrate his own plans for the future aggrandizement of France.

The end of the old order

The chief victims of the Final Recess were the free cities and the ecclesiastical territories. They fell by the dozens. Too weak to be useful allies of Napoleon, they were destroyed by the ambition of their French conquerors and by the greed of their German neighbours. They could still boast of their ancient history as sovereign members of the Holy Roman Empire, but their continued existence had become incompatible with the establishment of effective government in central Europe. The principal heirs to their holdings were the larger secondary states. To be sure, Napoleon could not keep Austria and Prussia from making some gains in the general scramble for territory that they had helped make possible. But he was especially solicitous for the welfare of those German rulers, most of them in the south, who were strong enough to be valuable vassals but not strong enough to be potential threats. Bavaria, Württemberg, Baden, Hesse-Darmstadt, and Nassau were the big winners in the competition for booty that had been the main object of the negotiations. Napoleon's strategy had been in the classic tradition of French diplomacy, the tradition of Richelieu and Mazarin. The princes had been pitted against the emperor to enhance the role that Paris could play in the affairs of central Europe. Yet neither the states that gained nor those that lost territory felt resentment at being used as pawns in a political game to promote the interests of a foreign power. Whatever objections they raised against the settlement of 1803 were based on expediency and opportunism. The most serious indictment of the old order was that in the hour of its imminent collapse none of the rulers attempted to defend it in the name of the general welfare of Germany.

The Final Recess was the next to the last act in the fall of the Holy Roman Empire. The end came three years later. In 1805 Austria joined the third coalition of Great Powers determined to reduce the preponderance of France (the War of the Third Coalition, 1805–07). The outcome was even more disastrous than its participation in the first and the second coalitions. Napoleon forced the main Habsburg army in Germany to surrender at Ulm (Oct. 17, 1805); then he descended on Vienna, occupying the proud capital of his enemy; and finally he inflicted a crushing defeat on the combined Russian and Austrian armies at Austerlitz in Moravia (Dec. 2, 1805). Before the year was out Francis II had been forced to sign the humiliating Treaty of Pressburg (December 26), which signified the end of the dominant role his dynasty had played in the affairs of central Europe. He had to surrender his possessions in western Germany to Württemberg and Baden, and the province of Tirol to Bavaria. Napoleon's strategy of playing off princely against imperial ambitions had proved a brilliant success. The rulers of the secondary states in the south had supported him in the war against Austria, and in the peace that ensued they were richly rewarded. Not only did they share in the booty seized from the Habsburgs but they also were permitted to absorb the remaining free cities, petty principalities, and ecclesiastical territories. Finally, in an assertion of the rights of full sovereignty, the rulers of Bavaria and Württemberg assumed the title of king, while the rulers of Baden and Hesse-Darmstadt contented themselves with the more modest rank of grand duke. The last vestiges of the imperial constitution had now been destroyed, and central Europe was ready to receive a new form of political organization reflecting the power relationship created by the force of arms.

The Confederation of the Rhine

In the summer of 1806, 16 of the secondary states, encouraged and prodded by Paris, announced that they were forming a separate association to be known as the Confederation of the Rhine. Archbishop Karl Theodor von Dalberg was to preside over the new union as the "prince primate," while future deliberations among the members were to establish a college of kings and a college of princes as common legislative bodies. There was even talk of a "fundamental statute" that would serve as the constitution of a rejuvenated Germany. Yet all these brave plans were never more than a facade for the harsh reality of alien hegemony in central Europe. Napoleon was proclaimed the "protector" of the Confederation of the Rhine, and a permanent alliance between the member states and the French Empire obliged the former to maintain substantial military forces for the purpose of mutual defense. There could be no doubt whose interests these troops would serve. The secondary rulers of Germany were expected to pay a handsome tribute to Paris for their newly acquired sham sovereignty. On August 1 the confederated states proclaimed their secession from the empire, and a week later, on Aug. 6, 1806, Francis II announced that he was laying down the imperial crown. The Holy Roman Empire thus came officially to an end after a history of a thousand years.

Period of French hegemony in Germany. The long conflict between emperors and princes in Germany had concluded with the triumph of the latter. Yet the victors soon discovered that, instead of achieving independence, they had merely exchanged one master for another. Indeed, they were more subordinate now to the wishes of Paris than they had been to those of Vienna. Napoleon gradually induced or forced all the states of Germany except Austria and Prussia, 36 states in all, to join the confederation. The inclusion of the two leading powers of central Europe might have proved troublesome for him or even dangerous. The participation of the secondary states, on the other hand, provided him with reliable mercenaries who owed him too much and feared him too much to oppose his wishes. He was free to do as he liked in the region between the Rhine and the Elbe. In order to enforce his embargo on continental trade with England, he annexed the entire coastline along the North Sea. When that was not enough, he added Lübeck on the Baltic to the French Empire. He carved out the kingdom of Westphalia for his brother Jérôme and the grand duchy of Berg for his brother-in-law Joachim Murat. He was the undisputed master of all of middle Germany.

After the formation of the Confederation of the Rhine, there was only one state in central Europe that had not yet been forced to submit to France. But the leaders of Prussia hesitated and wavered in their policy until they lost the opportunity of profiting from the War of the Third Coalition. Had they joined Austria and Russia against Napoleon, they might have kept him from gaining hegemony over Germany. Or had they become the allies of Napoleon, they might have established a sphere of influence in the region north of the Main River. As it was, they waited until they fell between two stools. They finally declared war against the French in October 1806, after Austria had been forced to surrender, Russia had decided to retreat, and the secondary states had become the vassals of Paris. Yet public opinion in the Prussian capital remained confident that the army of Frederick the Great would prove a match for the conqueror of Europe.

The subjection of Prussia

The result of such self-deception was a military disaster of unparalleled magnitude. In the two simultaneous battles of Jena and Auerstädt (Oct. 14, 1806) the Hohenzollern armies were completely routed, and the road to Berlin lay open before the French invaders. The city was occupied on October 27.

More disastrous than the military defeat, however, was the moral collapse of a state that had taught its citizens that obedience to authority was the supreme political virtue. The civilian population never thought of offering resistance to the advancing enemy. Even many army officers were so disheartened by Napoleon's success that they surrendered one fortified position after another without a fight. Frederick William III had to pay a terrible price for the policy of his ancestors, who had built efficient government at the expense of civic initiative. He tried to hold out in East Prussia, hoping that the Russian armies, which were still at war with Napoleon, would help him regain the rest of his kingdom. But when in July 1807 Alexander I concluded peace with France at Tilsit, the unfortunate Hohenzollern had no choice but to follow suit. The treaty that he was forced to sign was a catastrophe. Prussia lost almost half its territory and population, including most of the Polish possessions in the east as well as all of the territories west of the Elbe. Subsequent agreements, moreover, imposed a heavy indemnity, a military occupation, and a reduction in the size of the army. The proud monarchy of Frederick the Great had been reduced to a secondary state in Germany.

Central Europe remained under the dominant influence of France for more than a decade. That influence was at first limited and indirect, then pervasive and overpowering. Yet it was during this period of alien preponderance that Germany for the first time felt the stirrings of liberalism and nationalism. The regions that had become part of the French Empire experienced at first hand the advantages of efficient centralized government in which equality before the law and freedom of opportunity were accepted principles. Those states that retained a pseudo-independence as satellites of Napoleon, moreover, sought to imitate the example of their master, partly in order to gain his favour, partly in order to emulate his success. One government after another began to remove religious disabilities, relax economic restrictions, eliminate servile obligations, and centralize administrative functions. Above all, constitutional rule and popular representation ceased to seem Utopian to men of property and education who had witnessed the stirring events of the years since 1789. The French hegemony also led to the birth of nationalism in central Europe. For one thing, the achievement of political unity became a distinct possibility, once the territorial fragmentation of the Holy Roman Empire had come to an end. The presence of foreign invaders, furthermore—arrogant, overbearing, and avaricious—aroused among the Germans a sense of nationality that they had never felt in the tranquil days of the old order. Finally, an example of what great deeds the love of fatherland could inspire lay before all who admired or envied the triumphs of Napoleon. The ideal of cosmopolitan individualism that had been generally accepted in the 18th century began to give way before a growing consciousness of national identity. Yet the fact that the concepts of constitutional freedom and national unity were not indigenous but arose in response to foreign domination had an important effect on the form they assumed in central Europe.

Every German state felt the influence of the new principles of government and economy that the period of French hegemony had introduced, but nowhere was that influence more profound or fruitful than in Prussia. For only in the hour of deepest humiliation did the Hohenzollern kingdom finally make an effort to adapt its structure to the changing political and social conditions that it had stubbornly ignored during the years of greatness. Between 1806 and 1813 the statesmen in Berlin initiated a revolution from above in order to transform a rigid despotism into a popular monarchy supported by the loyalty of a free citizenry. Out of the disasters of Jena and Tilsit emerged a group of gifted reformers who sought to



Germany in 1807 after reconstruction by Napoleon.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London

The Prussian reformers: Stein and Hardenberg

prepare the way for the regeneration of their country. The leading figures in this movement for civic reconstruction were the civil servants Karl vom Stein and Karl August von Hardenberg and the military commanders Gerhard von Scharnhorst and August Neithardt von Gneisenau. Among their most important achievements was the abolition of serfdom, a measure designed to create citizens out of human beasts of burden. Yet, while it gave the peasant personal freedom, the government failed to provide him with economic independence. Most of the land remained in the hands of the aristocracy, which therefore continued to dominate the countryside politically as well as socially. More successful was the law establishing municipal self-government. Thereafter, the cities of the kingdom were to be administered by officials chosen not by the central bureaucracy but by the propertied inhabitants of the cities themselves. The autonomy of the cities, it was hoped, would help train a politically conscious and active middle class. The most effective reforms, however, were those introduced in the armed forces. After the officers who had shown themselves incompetent during the war were dismissed or retired, the high command carried out a thorough reorganization of the military system. Discipline became more humane, promotion was made a reward for merit, the method of recruitment was improved, and the training in tactics was modernized. Most important, the army's leaders sought to instill in the soldiers a new spirit rooted in inner conviction rather than unquestioning obedience. Defeat had changed Prussia from a garrison state into a centre of political and intellectual ferment.

Liberalism and nationalism became increasingly vehement in Germany as the burden of French domination grew progressively heavier. The financial sacrifices occasioned by the subordination to Napoleon reinforced the personal resentments aroused by his ruthless statecraft. Before long a network of secret organizations had sprung up in central Europe seeking the expulsion of the foreign invaders. Yet it would be a mistake to think that all Germans regarded the hegemony of France as an unmitigated evil. There were in fact wide differences of opinion among them. The rulers of the secondary states and their supporters in the army and the bureaucracy saw in Napoleon the instrumentality of their new importance. Many reformers in the south—the Bavarian statesman Maximilian von Montgelas, for example—believed that only French influence had made possible the modernization of government in Germany. Some publicists continued to argue, moreover, that the political disunity of central Europe was a natural result of its historic experience and reflected its essential character. To be sure, those who opposed alien domination outnumbered those who accepted it. But even among the former there was no agreement regarding the future political structure of the nation. Many of them dreamed of a liberal and united fatherland that would take its place among the great powers of Europe. Others were willing to settle for a loose association of governments, similar to the Confederation of the Rhine, which could safeguard the interests of the secondary states against Prussia and Austria. Still others hoped for a complete restoration of the old order in which they had grown up and to which they longed to return. And then there were the broad masses of the population of central Europe, exploited, illiterate, and uninformed. They remained by and large indifferent to the crosscurrents of political thought, seeking nothing more than an improvement in their standard of living and the preservation of their way of life. Germany was beginning to move toward new norms of civic and social value, but the transformation of political attitudes was gradual and intermittent.

That the growth of the ideals of unity and freedom was slow became apparent during the first serious effort to throw off the yoke of foreign domination in central Europe. The Austrian government concluded in 1809 that the reverses that Napoleon had been encountering in Spain presaged a general uprising against the French hegemony on the Continent. The result was an ill-fated attempt at a war of liberation, in which the Habsburg troops challenged Napoleon for the fourth time, only to go down in defeat once again. Appeals from Vienna to the people of Germany found little response except in Tirol and among a few nationalist hotspurs in the north. The princes refused to risk French wrath until they could be sure of ultimate victory, while their subjects refused to rise against French oppression without princely approval. The result was that the war in central Europe, unlike the one in the Iberian Peninsula, was waged primarily by regular forces rather than by guerrilla bands. Archduke Charles gained important successes for the Austrian army at Aspern and Essling (May 21–22, 1809), an indication that the strategic mastery of the French was drawing to a close. But at Wagram (July 5–6) Napoleon was able to work the last of his military miracles. Vienna had to sue for peace once more, the Treaty of Schönbrunn (October 14) ceding Salzburg to Bavaria, West Galicia to the grand duchy of Warsaw, and the Adriatic coastland to France. The defeat finally persuaded the emperor, who had exchanged the title Francis II of the Holy Roman Empire for Francis I of the Austrian Empire, that resistance would be as futile in the future as it had been in the past. He therefore adopted a policy of collaboration with France signalized by the marriage of his daughter Marie-Louise to Napoleon. Germany continued to languish in the grip of foreign domination.

The defeat of Austria

The wars of liberation. A new struggle for liberation opened three years later with the defeat of Napoleon's *grande armée* in Russia. As the tsarist armies began to cross their western frontiers in December 1812, the crucial question became what reception they would find among the rulers and the inhabitants of central Europe. The first state to cut its ties to Paris was Prussia. It was not the king, however, but one of his generals, Ludwig Yorck von Wartenburg, who decided on his own initiative to cooperate with the Russians. Only hesitatingly and fearfully did Frederick William III then agree in February 1813 to a war against France, although public opinion in his kingdom greeted the outbreak of the conflict with enthusiasm. The other rulers of central Europe refused initially to follow the Prussian example. The members of the Confederation of the Rhine were still convinced of Napoleon's invincibility, while Austria preferred to see the combatants exhaust each other to the point at which it could play the role of mediator and arbiter. The foreign minister in Vienna, Clemens Lothar von Metternich, was afraid that the hegemony of France in central Europe might be replaced by that of Russia. He tried, therefore, to pursue a strategy of armed neutrality, hoping that he could persuade the opposing sides to accept a compromise by which an equilibrium would be maintained between Alexander I and Napoleon. This plan failed because of the obstinacy of the latter, who feared that concessions in foreign affairs would weaken his control over internal politics in France. The upshot was that in August 1813 Austria entered the conflict on the side of Russia and Prussia, and the balance of military power shifted in favour of the anti-French coalition. The faith of the secondary states in Napoleon's star began to weaken, and Bavaria became the first member to secede from the Confederation of the Rhine (October 8). One great allied victory would now suffice to bring all of Germany into the struggle against France.

The anti-French alliance of Russia, Prussia, and Austria

That victory came on Oct. 16–19, 1813, at the Battle of Leipzig. After four days of bitter fighting, the French army was forced to retreat, and its domination of central Europe was finally at an end. Before the year was out, Napoleon had withdrawn across the Rhine. Of all his conquests in Germany, only the left bank was still under the effective control of Paris. The Confederation of the Rhine promptly collapsed, as its members rushed to go over to the winning side before it was too late. The Rhineland was also reconquered early in 1814, after the allies had launched their invasion of France. In the course of the spring the capture of Paris, the restoration of the Bourbons, and the conclusion of peace in the first Treaty of Paris (May 30) ended the war of liberation except for the episode of the Hundred Days, when Napoleon briefly returned to power and was ultimately and finally beaten at Waterloo. The western frontier of central Europe was to remain essentially the same as at the time of the initial outbreak of hostilities more than 20 years before. New

state boundaries within Germany would still have to be determined, to be sure, and the problem of a new political organization of the nation awaited the victorious statesmen, but the period of foreign hegemony was over at last. The rulers of central Europe, relying partly on the forces of innovation, partly on those of tradition, had succeeded in freeing themselves from alien domination. Now they had to decide what use they would make of their freedom. Would they create a new polity of unity and liberty, which many reformers demanded, or would they reestablish the old order of absolutism and particularism, which the conservatives advocated? As the statesmen began to gather in Vienna in the fall of 1814 to restore peace to a continent ravaged by two decades of war, they pondered the problem of devising an enduring form of government for Germany.

Results of the Congress of Vienna. The men who, in the nine months from September 1814 to June 1815, redrew the map of Europe were diplomats of the old school. Francis I and Prince von Metternich of Austria, Frederick William III and Prince von Hardenberg of Prussia, Alexander I of Russia, Viscount Castlereagh of England, Talleyrand of France, and the representatives of the secondary states were all intellectual heirs of the 18th century. They feared the principles of the French Revolution, they scorned the theories of democratic government, and they opposed the doctrines of national self-determination. But they recognized that the boundaries and governments of 1789 could not be restored without modification or compromise. There had been too many changes in attitudes and loyalties that the rigid dogmas of legitimism were powerless to undo. The task before the peacemakers was thus the establishment of a sound balance between necessary reform and valid tradition capable of preserving the tranquillity that Europe desperately needed. The decisions regarding Germany reached during the deliberations in Vienna followed a middle course between innovation and reaction, avoiding extreme fragmentation as well as rigid centralization. The Confederation of the Rhine was not maintained, but neither was the Holy Roman Empire restored. Although the reforms introduced during the period of foreign domination were partly revoked, the practices of enlightened despotism were not entirely reestablished. Despite the complaints of unbending legitimists and the dire predictions of disappointed reformers, the peacemakers succeeded in creating a new political order in central Europe that endured for half a century. The long years of war and unrest that had convulsed Europe during the era of the French Revolution and Napoleon were followed by even longer years of stability and tranquillity.

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1964); George Philip & Son Ltd., London



The German Confederation, 1815.

The Germany that emerged in 1815 from the Congress of Vienna included 39 states ranging in size from the two great powers, Austria and Prussia, through the minor kingdoms of Bavaria, Württemberg, Saxony, and Hannover; through smaller duchies such as Baden, Nassau, Oldenburg, and Hesse-Darmstadt; through tiny principalities such as Schaumburg-Lippe, Schwarzburg-Sondershausen, and Reuss-Schleiz-Gera; to the free cities of Hamburg, Bremen, Lübeck, and Frankfurt am Main. The new boundaries in central Europe bore little resemblance to the bewildering territorial mosaic that had been maintained under the Holy Roman Empire, but there were still many fragments, subdivisions, enclaves, and exclaves, too many for the taste of ardent nationalists. Yet the overall pattern of state frontiers represented a significant improvement over the chaotic patchwork of sovereignties and jurisdictions that had characterized the old order. The peacemakers not only created more integrated and viable political entities but also altered the role that these entities were to play in the affairs of the nation. Without design or even awareness on the part of Frederick William III, his kingdom of Prussia assumed a pivotal position in Germany. The victorious powers, on guard against a revival of French aggression, decided to make Berlin the defender of the western boundary of central Europe. The Rhineland and Westphalia, a region destined to develop into the greatest industrial centre on the Continent, became Hohenzollern provinces. More than that, the king agreed at the urging of Alexander I to cede the bulk of his Polish possessions to Russia in return for a substantial part of Saxony. Prussia, which at the end of the 18th century had been in the process of becoming a binational state, was thrust back into Germany and given a strategic position on both frontiers of the nation. The centre of gravity of Austria, on the other hand, shifted eastward. Francis I had decided to abandon the historic role of his state as protector of the Holy Roman Empire against the Bourbons for the sake of greater geographic compactness and military defensibility. The possessions in southern and western Germany were surrendered along with the Austrian Netherlands in return for Venetian territory on the Adriatic. The Habsburg empire thus became less German in composition and outlook as its focus shifted in the direction of Italy and the Balkans. The consequences of this territorial rearrangement were to be far-reaching.

Germanic Prussia and the polyglot Austrian Empire

THE AGE OF METTERNICH AND THE ERA OF UNIFICATION: 1815–71

Reform and reaction. In place of the Holy Roman Empire the peacemakers of the Congress of Vienna had established a new organization of states of central Europe, the German Confederation. This was a loose political association in which most of the rights of sovereignty remained in the hands of the member governments. There was no central executive or judiciary, only a federal Diet meeting in Frankfurt am Main to consider common legislation. The delegates who participated in its deliberations were representatives appointed by and responsible to the rulers whom they served. The confederation was in theory empowered to adopt measures strengthening the political and economic bonds of the nation. In fact it remained a stronghold of particularism, unwilling to sacrifice local autonomy in order to establish centralized authority. It was designed essentially to defend the interests of the secondary states and the Habsburgs. The former, jealously guarding the independence and importance they had gained during the period of French hegemony, were opposed to any reform that might limit their sovereignty. The latter believed that only a decentralized form of political union in Germany would give them enough freedom of action to pursue their non-German objectives. The confederation was thus from the outset an ally of localism and traditionalism. To the nationalists, whose hopes had

risen so high during the war of liberation, it seemed to be an instrument of blind reaction. Yet the truth is that the confederal system established in 1815 accurately reflected the slow development of civic consciousness and economic integration in central Europe. The militant reformers who demanded the centralization of government were a vocal but small minority. The lower classes accepted the territorial and constitutional decisions of the Congress of Vienna without a murmur of protest. The weakness of the peace settlement was not its failure to embody present realities but its inability to adjust to future changes. What had been a reasonable adaptation to the political needs of an agrarian and rural society became a hopeless anachronism 50 years later in the age of factories and railroads. This was the fatal flaw in the German Confederation.

Major flaw of the German Confederation

Yet the reform movement that had begun under the impact of the French hegemony did not end with the downfall of Napoleon. It continued to exert influence over affairs of state for another few years, before the forces of authoritarianism and particularism crushed it. That influence was strongest in southern Germany, where the political example of western Europe had made the deepest impression. There, many civil servants, court officials, army officers, and even aristocratic landowners came to believe that the future of the state depended on its readiness to reform civic institutions in accordance with liberal theories. In the years following Waterloo, one government in the south after another promulgated a constitution: Bavaria and Baden in 1818, Württemberg in 1819, and Hesse-Darmstadt in 1820. These constitutions established representative assemblies, elected by the propertied citizens, whose assent was required for the enactment of legislation. Their purpose was not only to win for the crown the support of the educated classes of society but also to engender a sense of unity in a heterogeneous population that still had diverse allegiances and traditions. To the north there were also persistent echoes of the reform movement.

The followers of Karl vom Stein were still influential in the councils of state, and Frederick William III of Prussia at first seriously considered ways of fulfilling the promise he had made in 1815 to establish constitutional government. The agitation for political reorganization was loudest, however, among university students, who formed patriotic groups known as *Burschenschaften.* They demanded the abandonment of the confederal system, the establishment of greater unity, and the achievement of national power. Gathering in 1817 at the Wartburg, a castle near Eisenach, they listened to veiled denunciations of the existing order and consigned to flames various symbols of traditional authority. The rulers of Germany began to stir uneasily at this bold display of defiance of legitimate government.

The chief strategist of the forces hostile to reform was Metternich. Not only did he reject the teachings of liberalism and nationalism in principle, but also, as the leading statesman of the Habsburg empire, he recognized that the establishment of centralized authority in Germany would seriously impede the policies his government was pursuing in Hungary, Italy, and the Balkans. When on March 23, 1819, an unbalanced student, Karl Ludwig Sand, assassinated the conservative playwright and publicist August von Kotzebue, Vienna persuaded the princes of the German Confederation that they were facing a dangerous attempt to overthrow the established order in central Europe. The result was a series of repressive measures called the Carlsbad Decrees, which the federal Diet adopted on Sept. 20, 1819. General censorship was introduced, and the *Burschenschaften* were outlawed. This first major success of the conservative counteroffensive had an important effect on the struggle within the state governments between the advocates and the opponents of reform. In Prussia the liberal members of the ministry were forced to resign, and the plan to promulgate a constitution for the kingdom was rejected. This shift to the right by Berlin encouraged authoritarian tendencies among the secondary states of the north, which soon abandoned their own constitutional projects. By the end of 1820 the reform movement, which had begun some 15 years before, came to a complete halt.

Repression and the Carlsbad Decrees

It had succeeded in altering the political and economic structure of society, but it had been unable to establish a tradition of liberal government and national loyalty in central Europe. The forces of particularism and legitimism, deriving their chief support from the landowning nobility and the conservative peasantry, remained strong. The foundation of bourgeois civic consciousness and material prosperity on which England and France had built their representative institutions was still lacking beyond the Rhine. The ideal of free government was introduced in Germany not as the fruit of an industrial or a political revolution but as the imitation of a foreign example and the reaction against a foreign oppression.

The established order was once again threatened briefly in the wake of the July days of 1830 in France. The news that there had been a successful insurrection against the Bourbons in Paris had an electrifying effect throughout the Continent. In central Europe there were sympathetic uprisings in some of the secondary states of the north. The rulers of Brunswick, Saxony, Hannover, and Hesse-Kassel, seeking to forestall more extreme demands, agreed to promulgate liberal constitutions. A mass meeting of southern radicals at Hambach Castle in the Palatinate (May 1832), moreover, expressed its approval of national unification, republican government, and popular sovereignty. A group of militant students even launched a foolhardy attempt to seize the city of Frankfurt am Main, dissolve the federal Diet, and proclaim a German republic. The effect of such harebrained schemes was predictable. As the princes of the German Confederation gradually recovered from their initial fear of the revolutionary movement, they began to oppose with increasing vigour plans to alter the existing system of government. Again, Metternich took the lead in the effort to crush liberalism and nationalism. Under his direction the federal Diet adopted additional repressive measures reinforcing the position of the crown in state politics, limiting the power of the legislature, restricting the right of assembly, enlarging the authority of the police, and intensifying the censorship. Within a few years the opposition had been subdued, and the German Confederation could continue to vegetate in its cozy provincialism. Not until the middle years of the century did a new and more violent outburst of political disaffection shake the foundations of the system of authority that the Congress of Vienna had erected.

Effects of the July Revolution of 1830

Evolution of parties and ideologies. Although the critics of the established order could be defeated, they could not be silenced. The struggle between the supporters and the adversaries of the existing form of government led to the emergence of a rudimentary party system in the German Confederation. In the legislative assemblies of the secondary states the proponents of reform began to meet, plan, organize, and propagandize. The defenders of legitimism were thereby forced in turn to concert their strategy and to publicize their program. Even in Prussia and Austria, where there were as yet no constitutions or parliaments, political criticism could be expressed obliquely through clubs, meetings, newspapers, pamphlets, and petitions. The result was the gradual development of amorphous civic associations held together by common convictions regarding the nature of state and society. These primitive groupings were only the raw material out of which disciplined political parties were slowly fashioned in the course of the century. They still lacked a clear sense of purpose and the systematic propagation of belief characteristic of a fully mature system of parliamentary politics. Yet they became the instrumentalities by which disaffected groups in the community could express their opposition to the established order. They reflected the fact that the civic attitudes of the period of the Restoration were no longer those of the age of enlightened despotism. There were men in central Europe now who refused to submit without question to princely authority, seeking freedom only in the inner recesses of the soul. The change in the form of economy and the structure of society produced by the beginnings of industrialization led to an alternation in the system and organization of politics.

The most important opponents of legitimism and particularism were the liberals or moderates. Deriving their sup-

The liberals or moderates

port primarily from industrialists, merchants, financiers, mineowners, railroad promoters, civil servants, and university professors, they represented the opposition of the well-to-do bourgeoisie to a form of government in which an aristocracy of birth rather than of talent predominated. Their political principles favoured a monarchical system of authority, but the crown was to share its powers with a parliament elected by the propertied classes. Influence in public affairs should be accessible to all citizens who had demonstrated through the acquisition of wealth and education that they were capable of exercising the franchise intelligently. While the liberals resented the inherited privileges of the nobility, they also feared the proletariat. The man who lived in poverty and ignorance, they reasoned, was ripe for demagoguery and insurrection. The path of civic wisdom was therefore the *juste-milieu* between royal absolutism and mob rule, which had been established in England by the Reform Bill of 1832 and in France by the regime of Louis-Philippe. In economics, the teachings of liberalism advocated a policy of unrestrained competition by which wealth would become the reward of business acumen rather than the perquisite of corporative privilege. Guild monopolies, government regulations, and entailed estates were to be abolished as violations of that freedom of enterprise that alone could ensure the well-being of all society. The liberals favoured the transformation of the German Confederation into a national monarchy in which the states' rights would be curtailed but not destroyed by a central government and a federal parliament.

The democrats or radicals

Farther to the left stood the democrats or radicals, whose following was made up largely of small businessmen, petty shopkeepers, skilled workers, independent farmers, publicists, journalists, lawyers, and physicians. They looked with scorn on the golden mean between autocracy and anarchy that the liberals were seeking. They preferred an egalitarian form of authority in which not parliamentary plutocracy but popular sovereignty would be the underlying principle of government. Their supporters drew inspiration from the French rather than the English or the American Revolution. Their ideal of petty bourgeois democracy was the Jacobin republic of 1793 as an instrument to shape the energies and aspirations of the people into a disciplined force for political and social reform. The spokesmen of this ideal could not openly demand the overthrow of monarchical institutions without risking imprisonment. Yet, while they were forced to accept the crown as a political institution, they sought to transfer its power to a parliament elected by equal manhood suffrage. The masses would thereby become the ultimate arbiter of politics. The democrats were also willing to accept government regulation of business activity as a means of improving the economic position of the lower classes, although their belief in the sanctity of private property was as firm as that of the liberals. In their advocacy of national unification, however, they were less solicitous about royal prerogatives and state rights. While not as numerous as the moderates, the radicals remained an important source of opposition to the established order.

The growth of criticism directed against the political system of the Restoration forced its supporters to define their ideological position with greater precision. The old theories of monarchy by divine right or despotic benevolence offered little protection against the assaults of liberalism and democracy. Legitimism, whose defenders came mostly from the landed nobility, the court aristocracy, the officer corps, the upper bureaucracy, and the established church, began therefore to advance new arguments based on conservative assumptions about the nature of man and society. The relationship between the individual and government, so the reasoning went, cannot be determined by paper constitutions founded on a doctrinaire individualism. Human actions are not motivated solely by rational considerations but by habit, feeling, instinct, and tradition as well. The impractical theories of visionary reformers fail to take into account the historic forces of organic development by which the past and the present shape the future. To assert that all men are equal is to ignore differences in rights and duties expressing differences in birth, class, background, education, and tradition. The dogmas of constitutional authority and parliamentary government are merely a facade behind which a self-seeking bourgeoisie seeks to disguise its lust for power. An enduring form of government can be built only on the traditional institutions of society: the throne, the church, the nobility, and the army. Only a system of authority legitimated by law and history can protect the worker against exploitation, the believer against godlessness, and the citizen against revolution. According to these tenets, the political institutions of the German Confederation were valid, because they represented fundamental convictions deeply embedded in the spirit of the nation.

The conservatives or legitimists

Economic changes and the Zollverein. The struggle of parties and ideologies during the Restoration reflected far-reaching changes in the structure of the economy and the community. The most significant of these changes was the rise of large-scale industry in central Europe. Techniques of mechanization, introduced in textile mills and coal mines, spread to other branches of manufacture and exerted pressures that influenced the entire economic life of the nation. The transportation network improved with the construction of railroads, steamships, bigger highways, and better canals. Banking institutions and private investors began to transfer their funds from government bonds and commercial ventures to manufacturing enterprises. Millowners, ironmasters, railroaders, financiers, and stockbrokers gradually formed a new middle class whose wealth was derived primarily from industrial activity and whose growing economic importance encouraged among its members a demand for greater political influence. Skilled handicraftsmen, constituting the bulk of the urban working class, could not compete successfully with the factories. The conflict between industrial and preindustrial forms of output was aggravated, moreover, by important demographic changes in the period following the Congress of Vienna. Population began to shift from country to city, although a majority of the inhabitants of the German Confederation continued to live in rural communities.

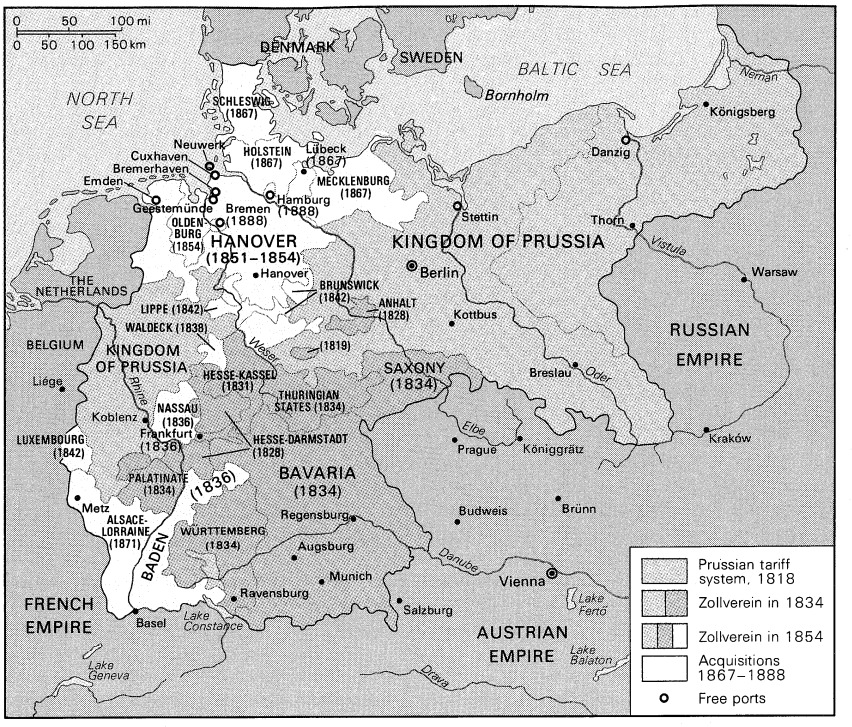
The growth of large-scale industry

Agriculture also went through as difficult a period of reorganization and rationalization as industry. The end of serfdom had led in the regions east of the Elbe to the expansion of large estates belonging to aristocratic landowners but cultivated by a propertyless rural proletariat. Peasant emancipation in Prussia allowed the Junkers to enlarge their lands by absorbing the holdings of small farmers. The result was the continuing economic, social, and political domination of the village by the nobility in the eastern provinces of the Hohenzollern kingdom. The squirearchy entrenched in Pomerania, Brandenburg, Silesia, and East Prussia controlled agriculture, commanded the army, directed the bureaucracy, and influenced the court. It constituted a powerful force for conservatism and particularism.

The reorganization of agriculture

West of the Elbe the basic problem was not landlessness but overpopulation. The aristocracy along the Rhine and the Danube was often willing to give the peasantry possession of the soil in return for a substantial payment. The farmer was thereby saddled with heavy financial obligations. Many rustics tried to escape poverty by emigrating to the New World; those who remained faced swift demographic expansion and often had to subdivide small holdings until they yielded no profit. Civil discontent mounted among impoverished villagers who lacked employment in industry.

The system of authority in the German Confederation was thus being undermined by the struggle of artisans against industrial mechanization, by the disaffection of peasants hungry for land, and above all by the criticism of businessmen groaning under the shackles of particularism. Industrialists and financiers had to overcome the barriers created by a variety of monetary systems, commercial regulations, excise taxes, and state boundaries. It is little wonder that the bourgeoisie of central Europe turned increasingly to the teachings of liberalism and nationalism. Yet the established order did make a major attempt to meet the needs of the business community. Before long, several of the more important secondary governments concluded agreements with Berlin by which a sizable free



The growth of the German Zollverein.

trade area was established in the heart of central Europe.

The Zollverein

In 1834 the Zollverein, or Customs Union, including most of the states of the German Confederation, came into existence. Only Austria and the northwest coastland remained aloof. The industrialists of the Habsburg empire, who wanted their products protected against outside competition, felt that the tariff rates of the new association were too low for their needs, whereas the merchants and bankers of the coastal region, who depended on imports, thought that they were too high. Yet for some 25 million Germans the Zollverein meant in effect the achievement of commercial unification without the aid of political unification. The Prussian government, moreover, acquired a powerful new weapon in the struggle against Austria for the dominant position in central Europe. Still, although the Zollverein helped meet the most pressing demands of the middle class for economic consolidation, it could not surmount all the material disadvantages of a particularistic form of government. People of means and education continued to grumble about the confederal system under which they lived, while the masses became increasingly restless under the pressure of social dislocation.

The revolutions of 1848–49. The hard times that swept over the Continent in the late 1840s transformed widespread popular discontent in the German Confederation into a full-blown revolution. After the middle of the decade a severe business depression halted industrial expansion and aggravated urban unemployment. At the same time, serious crop failures led to a major famine in the entire area from the Irish Sea to Russian Poland. In central Europe the hungry '40s drove the lower classes, which had long been suffering from the economic effects of industrial and agricultural rationalization, to the point of open rebellion. There were sporadic hunger riots and violent disturbances in several of the states, but the signal for a concerted uprising did not come until early in 1848 with the exciting news that the regime of the bourgeois king Louis-Philippe had been overthrown by an insurrection in Paris (February 22–24). The result was a series of sympathetic revolutions against the governments of the German Confederation, most of them mild but a few, as in the case of the fighting in Berlin, bitter and bloody. When on March 13 Metternich, the proud symbol of the established order, was forced to resign his position in the Austrian Cabinet, the princes hastened to make peace with the opposition in order to forestall republican and socialist experiments like those in France. Prominent liberals were appointed to the state ministries, and civic reforms were introduced safeguarding the rights of the citizens and the powers of the legislature. But even more important was the attempt to achieve political unification through a national assembly representing all of Germany. Elections were held soon after the spring uprising had subsided, and on May 18 the Frankfurt National Assembly met to prepare the constitution for a free and united fatherland. Its convocation represented the realization of the hopes that nationalists had cherished for more than a generation. Within the space of a few weeks, those who had fought against the particularistic system of the Restoration for so long suddenly found themselves in power with a popular mandate to rebuild the foundations of political and social life in central Europe. It was an intoxicating moment.

The Frankfurt National Assembly

The forces that had defeated the establishment order, however, soon discovered that they were in disagreement regarding the use to be made of their common victory. There were, first of all, sharp differences between the liberals and the democrats. While the former had comfortable majorities in most of the state legislatures as well as in the Frankfurt parliament, the latter continued to plead, agitate, and conspire for a more radical course of action. There were also bitter disputes over the form that national unification should assume. The Grossdeutsch (Great German) movement maintained that only Austria, the state whose rulers had worn the crown of the Holy Roman Empire for 400 years, could successfully guide the united fatherland. The Kleindeutsch (Little German) party, on the other hand, argued that the Habsburgs had too many Slavic, Magyar, and Italian interests to work single-mindedly for the greatness of Germany. The natural leader of the nation was Prussia, whose political vigour and geographic position would provide efficient government and military security for central Europe. Finally, there was a basic conflict between the proletariat, which wanted protection against mechanized production and rural impoverishment, and the bourgeoisie, which sought to use the political power it had gained in order to promote an industrial capitalism based on freedom of enterprise. Once the spring uprising was over, the parties and classes that had participated in it began to quarrel about the nature of the new order that was to take the place of the old. Popular support for the revolution, which had made the defeat of legitimism during the March days possible, began to dwindle, discouraged by the realization that the liberals would do no more to solve the problems of the masses

than the conservatives had done. While the Frankfurt parliament was debating the constitution under which Germany would be governed, its following diminished and its authority declined. The forces of the right, recovering from the demoralization into which they had been plunged by the revolution, began to plan a counterrevolution.

Their first major victory came in Austria, where the young emperor Francis Joseph found an able successor to Metternich in his prime minister Prince Felix zu Schwarzenberg. In the course of the summer of 1848 the Habsburg armies crushed the uprising in Bohemia and checked the insurrection in Italy. By the end of October they subjugated Vienna itself, the centre of the revolutionary movement, and now only Hungary was still in arms against the imperial government. At the same time, in Prussia the irresolute Frederick William IV had been gradually persuaded by the conservatives to embark on a course of piecemeal reaction. Early in December he dissolved the constituent assembly that had been meeting in Berlin, promulgated by his own authority a middle-of-the-road constitution for the kingdom, and proceeded little by little to reassert the prerogatives of the crown. Among the secondary states there was also a noticeable shift to the right, as particularist princes and legitimist aristocrats began to regain their courage. By the time the Frankfurt parliament completed its deliberations in the spring of 1849, the revolution was everywhere at ebb tide. The constitution that the National Assembly had drafted provided for the creation of a federal union at the head of which was to stand a hereditary emperor with powers limited by a popularly elected legislature. Since the Austrian government had already indicated that it would oppose the establishment of a centralized system in Germany, the imperial crown was offered to the king of Prussia. Frederick William IV hesitated, brooded, and agonized, but in the end he refused the dignity whose authority was in his opinion too restricted. This rejection of political consolidation under a liberal constitution destroyed the last chance of the revolutionary movement for success. The moderates, admitting failure, went home to mourn the defeat of their hopes and labours. The radicals, on the other hand, sought to attain their objectives by inciting a new wave of insurrections. Their appeals for a mass uprising, however, were answered mostly by visionary intellectuals, enthusiastic students, radical politicians, and professional revolutionaries. The lower classes remained by and large indifferent. There was sporadic violence, especially in the southwest, but troops loyal to princely authority had little difficulty in defeating petty bourgeois democracy. By the summer of 1849 the revolution, which had begun a year earlier amid such extravagant expectations, was completely crushed.

Suppression of the revolutions

The 1850s: years of political reaction and economic growth. The attempt to achieve national unification through liberal reform was followed by an attempt to achieve it through conservative statesmanship. Frederick William IV had refused to accept an imperial crown vitiated by parliamentary government, but he was willing to become the head of a national federation in which the royal prerogative remained unimpaired. While the Austrian armies were still engaged in the campaign against the revolution in Hungary, Berlin began to exert diplomatic pressure on the secondary governments to join in the formation of a new federal league known as the Prussian Union. If Frederick William IV had acted with enough determination, he might have been able to reach his goal before Francis Joseph could intervene effectively in the affairs of Germany. But he allowed his opportunity to slip away. Though he succeeded through threats and promises in persuading most of the princes to accept his proposals, no irrevocable commitments had been made by the time the Hungarians were defeated in August 1849. Vienna could now proceed to woo the secondary governments, which had in most cases submitted to Prussia only out of weakness and fear. Basically they remained opposed to sacrificing their sovereignty in order to exalt the Hohenzollern dynasty. When Schwarzenberg suggested the reestablishment of the old federal Diet, he won the support of many rulers who had agreed to follow Berlin against their will. The nation was now divided into two camps, the Prussian Union on one side and the revived German Confederation on the other. It was only a question of time before they would clash. When both Austria and Prussia decided to intervene in Hesse-Kassel, where there was a conflict between the supporters and the opponents of the prince, Germany stood on the brink of civil war. But Frederick William IV decided at the last moment to back down. His fear overcame his pride, especially after Nicholas I of Russia indicated that he supported Vienna in the controversy. By the Punctation of Olmütz of Nov. 29, 1850, the Prussians agreed to the restoration of the German Confederation, and the old order was fully reestablished in all its weakness and inadequacy.

Revival of the German Confederation

The years that followed were a period of unmitigated reaction. Those who had dared to defy royal authority were forced to pay the penalty of harassment, exile, imprisonment, or even death. Many of the political concessions made earlier, under the pressure of popular turmoil, were now restricted or abrogated. In Austria, for example, the constitution that had been promulgated in 1849 was revoked, and legitimacy, centralization, and clericalism became the guiding principles of government. While in Prussia the constitution granted by the king remained in force, its effectiveness was reduced through the introduction of a complicated system of election by which the ballots were weighted in accordance with the income of the voters. The consequence was that well-to-do conservatives controlled the legislature. The secondary states returned to the policies of legitimism and particularism that they had pursued before the revolution. In Frankfurt am Main, where the federal Diet now resumed its sessions, diplomats continued to guard the prerogatives of princely authority and state sovereignty. The restoration of the confederal system also served the interests of the Habsburgs, who stood at the pinnacle of their prestige as the saviours of the established order. In Berlin, on the other hand, the prevailing mood was one of confusion and discouragement. The king, increasingly gloomy and withdrawn, came under the influence of ultraconservative advisers who preached legitimism in politics and orthodoxy in religion. The government, smarting under the humiliation suffered at the hands of Austria, was as timid in foreign as it was oppressive in domestic affairs. The people, tired of insurrection and cowed by repression, were politically apathetic. The German Confederation as a whole, rigid and unyielding, remained during these last years of its existence blind to the need for reform that the revolution had made clear.

Yet the 1850s, so barren in politics, were of the highest importance in economics, for it was during this period that the great breakthrough of industrial capitalism occurred in central Europe. The national energies, frustrated in the effort to achieve civic reform, turned to the attainment of material progress. The victory of the reaction was followed by an economic expansion as the business community began to recover from its fear of mob violence and social upheaval. The influx of gold from America and Australia, moreover, generated an inflationary tendency, which in turn encouraged a speculative boom. Not only did the value of industrial production and foreign trade in the Zollverein more than double in the course of the decade, but also new investment banks based on the joint-stock principle were founded to provide risk capital for factories and railroads. The bubble burst in 1857 in a financial crash that affected the entire Continent. For many investors the price of overoptimism and speculation turned out to be misfortune and bankruptcy. Yet Germany had now crossed the dividing line between a preindustrial and an industrial form of economy. Although the rural population still outnumbered the urban, the tendency toward industrialization and urbanization had become irreversible. And this in turn had a profound effect on the direction of politics. For, as wealth continued to shift from farming to manufacturing, from the country to the city, and from the aristocracy to the bourgeoisie, the pressure for a redistribution of political power also gained strength. While the reactionaries were solemnly proclaiming the sanctity of traditional institutions, economic change was undermining the foundation of those institutions. By the

Industrial capitalism

end of the decade a new struggle between the forces of liberalism and conservatism was in the making.

The 1860s and the triumphs of Bismarck. The revival of the movement for liberal reform and national unification at the end of the 1850s came to be known as the "new era." Its coming was heralded by scattered but distinct indications that the days of the reaction were numbered. In 1859 the defeat of Austria in the war against France and Piedmont had a profound effect on central Europe. For one thing, the maintenance of the authoritarian regime in Vienna depended on respect for its military strength. Now that the Habsburg armies had shown themselves to be vulnerable, popular unrest in the empire began to increase. Since autocracy was no longer an effective principle of government, Francis Joseph decided to experiment with a parliamentary form of authority. On Oct. 20, 1860, he promulgated a constitution (the October Diploma) for his domains, setting up a bicameral legislature with an electoral system favouring the bourgeoisie, and Austria ceased to be an absolutist state. The achievement of political consolidation in Italy, moreover, aroused hope and envy north of the Alps. If the Italians could overcome the obstacles of conservatism and particularism, why not the Germans? National sentiment in central Europe, dormant since the revolution, suddenly awoke. Patriotic organizations like the Nationalverein (National Union) and the Reformverein (Reform Union) initiated an agitation for a new federal union, the former advocating Prussian, the latter Austrian, leadership. Liberal publicists and politicians began to advance plans for the reconstruction of the German Confederation. Some of the states, detecting a shift in the trend of public opinion, decided to change their course accordingly. Here and there the conservative ministers of the reaction were retired or dismissed, and their place was taken by statesmen with more moderate views.

The most significant portent of a new age in politics, however, appeared in Prussia. In 1857 Frederick William IV, crushed by memories of the mass insurrections and diplomatic defeats that he had been forced to endure, suffered a mental breakdown. A year later his brother became regent, and the government in Berlin immediately began altering the direction of its policy. Prince William, although a man of conservative inclination, had little sympathy with the mystical visions and pious dogmas prevailing at the court during the period of reaction. He dismissed the Cabinet that had served his predecessor, announced a program of cautious reform in Prussian as well as German affairs, and won a popular endorsement of his course in the elections that gave the liberals control of the legislature. After a long period of discouragement, the advocates of civic reconstruction could once again look to the future with hope and expectation.

Yet there was an important difference between the political attitude of the liberals in 1858–59 from that of 1848–49. They came to feel during the new era that their defeat 10 years before had been due to an excess of idealism and exuberance. The fatal mistake of the revolution, they reasoned, had been the assumption that enthusiasm and selflessness could be translated into power and substituted for statesmanship. Now a more calculating policy, a policy of realpolitik, must be adopted. Not theory and rhetoric but negotiation and compromise would lead to the attainment of unity and freedom. The liberals therefore pursued at first a strategy of conciliation, anxious not to frighten the established order into blind resistance against all reform. In Prussia, for example, they waited patiently for the regent to move against the forces of disunity and oppression, confident that if they only gave him enough time he would obtain for them by royal authority what they could not seize through revolutionary violence. Yet it gradually became apparent that their hopes would not be realized. Prince William, who in 1861 became king in his own right, was a moderate conservative but a conservative nevertheless. As the advocates of reform grew increasingly restless, the more militant among them formed the Progressive Party, which sought to hasten the introduction of liberal legislation by exerting pressure on the government. The monarch, afraid that he was being pushed farther to the left than he wanted to go, became more adamant and uncompromising. Sooner or later a conflict between crown and parliament was bound to arise.

The Prussian policy of realpolitik

It came in connection with the question of army reform, although, if that issue had not developed, there would undoubtedly have been another. The king wanted to strengthen the armed forces by increasing the number of line regiments and decreasing the size of the popular militia. The legislature, reluctant to enhance the power of the conservative officer corps, demanded a modification of the plan. The king refused, convinced that the politicians were attempting to gain control of the army, which he considered subject only to royal authority. The legislature responded by withholding approval of the budget to pay for the cost of army reform. A complete deadlock ensued. In the spring of 1862 the liberal ministers who had been appointed after the establishment of the regency were dismissed, and a conservative Cabinet took office. But the new leaders of the government were as unsuccessful as the old in resolving the crisis. William I began to think about abdicating in favour of his son, who was believed to have political views close to those of the parliamentary opposition. He was persuaded, however, to consider first the possibility of naming a new ministry led by Otto von Bismarck, the Prussian ambassador to Paris. There was a momentous interview between the monarch and the envoy, as a result of which the former abandoned all thought of retirement, and the latter became head of a Cabinet pledged to continue the struggle against the legislature. The battle between crown and parliament, which the liberals had been on the point of winning, was now to be waged without regard for constitutional provisions concerning the budget. On Sept. 23, 1862, the nation was startled by the news that a statesman with a reputation for unyielding conservatism had become the prime minister of Prussia. It meant that the established order, having successfully defended its interests against the forces of reform after 1815 and after 1848, was determined to fight to the bitter end against the new challenge to its predominance.

Advent of Otto von Bismarck

The constitutional conflict in the Hohenzollern kingdom continued for another four years. The legislature refused to approve the budget until its wishes concerning military reform had been met. Bismarck's government, after carrying out the controversial reorganization of the army, continued to collect taxes and disburse funds without regard for parliamentary authorization. The liberals condemned the prime minister as a violator of the law, while the prime minister denounced the liberals as blindly doctrinaire. Although the electorate remained on the side of the opposition, the Cabinet declared that it would not be swayed by party politics or parliamentary majorities. The broad masses of the population, it maintained, were still loyal to the crown. And so the struggle went on without prospect of alleviation or resolution. There were even dark prophecies of a violent uprising against a regime that was so indifferent to its constitutional obligations. Yet in fact Bismarck was not blind to the need for a reconciliation between crown and bourgeoisie. Despite his reputation as a fire-eating legitimist, he had a supple mind and recognized that the political principles of Frederick the Great could not solve the problems facing William I. He hoped, therefore, for an eventual reconciliation between the government and the legislature, but a reconciliation in which the prerogative of the monarch and the influence of the nobility would remain undiminished.

What Bismarck sought in essence was an alteration in the form of government that would create a facade of parliamentary institutions disguising the continuation of authoritarian policies. The middle class wanted to end the domination of the traditional forces in society, he calculated, but it also wanted to achieve national unification in central Europe. Here was the key to a solution of the constitutional conflict. Unity could be used to restrict freedom; nationalism could become the means of taming liberalism. Bismarck had concluded that the political integration of Germany was, in the long run, inevitable. If the established order did not effect it, the reformers, democrats, and revolutionaries would. Thus, it was in the interest of conservatism to take the task of centralization in hand, bring it to a successful conclusion, and create a

new system of authority compatible with the preservation of royal and aristocratic preponderance. Such a policy would make possible a compromise between crown and bourgeoisie by which the latter obtained the benefits of economic consolidation while the former retained the advantages of political domination. Through this strategy the prime minister hoped to end the constitutional conflict.

The defeat of Austria. The international situation was favourable to a program of unification in the German Confederation. Since its defeat in the Crimean War (1853–56), Russia had ceased to play a decisive role in the affairs of the Continent. England remained preoccupied with the problems of domestic reform. And Napoleon III was not unwilling to see a civil war east of the Rhine that he might eventually use to enlarge the boundaries of France. Bismarck could thus prepare for a struggle against Austria without the imminent danger of foreign intervention that had faced Frederick William IV. His first great opportunity came in connection with the duchies of Schleswig and Holstein, which were ruled by the king of Denmark but which were politically and ethnically tied to Germany. When the government in Copenhagen sought to make Schleswig an integral part of the Danish state in 1863, nationalist sentiment in central Europe was outraged. William I proposed to Francis Joseph that the two leading powers of the German Confederation should occupy the duchies in order to prevent the violation of an international agreement that had guaranteed their separate status. Afraid to let the Prussians act on their own, the emperor agreed, and in 1864 there was a brief war against Denmark that demonstrated the strength of the reorganized army of the Hohenzollerns. Danish hopes for foreign assistance proved illusory, and by the Peace of Vienna (October 30) the duchies became the joint possession of Prussia and Austria.

The Danish War and Schleswig-Holstein

The easy victory of the allies, however, was only the prelude to a bitter conflict between them. Vienna would have liked to see Schleswig-Holstein become an independent secondary state in the German Confederation, committed to a policy of particularism. Berlin, on the other hand, hoped for the outright annexation of the duchies or at least the indirect control of their government. Even more important than the disposition of the spoils of war, however, was the mounting rivalry between the two great powers for hegemony in central Europe. Bismarck would probably have been willing to collaborate with Austria in the division of Germany into two spheres of influence, a northern under the Hohenzollerns and a southern under the Habsburgs, but Francis Joseph was reluctant to restrict his authority and ambition to the region below the Main River. The result was a steady deterioration of relations between Vienna and Berlin. In 1865 their differences were papered over by the Convention of Gastein, which placed Schleswig under Prussian and Holstein under Austrian administration but which also reaffirmed the joint sovereignty of the two governments over the duchies. Still, this was only a temporary solution, and before long the danger of civil war in the German Confederation began to grow once again.

In the course of the spring of 1866 both sides stepped up their preparations for a military solution to the Austro-Prussian dualism. Bismarck concluded an alliance with Italy by which the latter was to receive Venetia as a reward for participating in a war against the Habsburg empire. He also sought to gain the support of public opinion in the German Confederation by introducing a motion in the federal Diet for the convocation of a national parliament elected by equal manhood suffrage. The Austrians in the meantime secured a promise of French neutrality in the event of hostilities and tried to win the adherence of the secondary states in the impending struggle. The last desperate attempts to preserve peace collapsed in June. Vienna announced that it would submit the question of the duchies to the federal Diet. Berlin, condemning this step as a violation of the Convention of Gastein, ordered its troops in Schleswig to expel the Austrians from Holstein. Francis Joseph in reply called on the other states of the confederation to mobilize their armies against the Prussian threat to domestic tranquillity, and central Europe trembled on the verge of civil war. The only question now was what the position of the secondary states would be. Most of them lined up behind Austria, which they regarded as the defender of their independence against the ambitions of Berlin. Bismarck's attempt to enlist the aid of the national movement by advocating reform of the confederal system thus failed. It alarmed the particularists without propitiating the centralists. Public opinion remained frightened and confused, distrusting one side and fearing the other. The future of the nation was decided not by popular insurrections or parliamentary deliberations but by the force of arms.

The Seven Weeks' War

The Seven Weeks' War between Prussia and Austria (June–August) produced a diplomatic revolution in Europe, destroying the balance of power that had been established 50 years before by the Congress of Vienna. Yet this momentous alteration in the international equilibrium was accomplished so swiftly that foreign diplomats had barely begun to grasp its implications before the struggle for hegemony in Germany ended. The Hohenzollern armies had a brilliant strategist in Helmuth von Moltke and a deadly weapon in the breech-loading needle gun. The Austrian high command, on the other hand, became irresolute and demoralized before a decisive encounter had even taken place. The Prussians succeeded in dividing and defeating the forces of the secondary states, and on July 3 they routed the Habsburg troops as well at the Battle of Königgrätz (Sadowa). The war was thus decided within a few weeks after its outbreak. Bismarck, refusing to be dazzled by the brilliance of the victory, urged the swift conclusion of an honourable peace. Not only did he feel that the preservation of a strong Austria was essential for the maintenance of stability on the Continent, but he also feared that a prolongation of hostilities would enable Napoleon III to intervene in the affairs of Germany. By the preliminary Peace of Nikolsburg (July 26) and the definitive Treaty of Prague (August 23), Francis Joseph was permitted to retain all of his possessions except Venetia, which had been promised to the Italians. There was to be no occupation and only a modest indemnity. The emperor had to acquiesce, however, in the Prussian annexation of Hanover, Nassau, Hesse-Kassel, Schleswig-Holstein, and Frankfurt am Main, in the dissolution of the German Confederation, and in the formation under Hohenzollern leadership of a new federal union north of the Main River. The contest between Berlin and Vienna that had determined the history of central Europe for more than a century was now over.

Bismarck's national policies: the restriction of liberalism. Bismarck's triumph in the military struggle led directly to his victory in the constitutional conflict. Before the outbreak of hostilities he had tried to reach an understanding with the liberal opposition, but the liberals hesitated to make peace with a statesman who had so flagrantly violated the fundamental law of the kingdom. The defeat of Austria changed all that. While the war was still in progress, general elections resulted in important gains for the right. Many voters, elated over the successes of the Prussian armies, expressed their confidence in the government by supporting its adherents at the polls. Some of the ultraconservatives hoped that the Cabinet would now capitalize on its triumph by suspending the constitution and establishing an authoritarian regime. Yet the prime minister recognized that such reactionary schemes would prove futile in the long run. What he wanted was not the suppression of liberalism but an accommodation with it. As soon as peace was concluded, he introduced in the legislature a bill of indemnity granting the government retroactive approval for its conduct of affairs of state without a legal budget. The consequence, as Bismarck had foreseen, was a split in the ranks of his adversaries. Those who argued that there could be no compromise with the principle of constitutional government rejected a reconciliation between crown and parliament based on mutual concessions, but many members of the opposition, who eventually formed the National Liberal Party, decided to accept the settlement offered by the prime minister. Their reasoning was that an obstinate resistance against the Cabinet would only condemn them to sterile dogma-

tism, whereas a willingness to accept what could not be prevented would enable them to influence official policy in the direction of greater freedom. This view prevailed, and on Sept. 3, 1866, the legislature approved the Bill of Indemnity, 230 to 75. By dividing the forces of reform and weakening their sense of purpose, Bismarck won as important a success in domestic affairs as the victory on the field of battle.

His subjugation of liberalism was made possible by the triumph of nationalism. It had long been the prime minister's belief that the achievement of unity could appease the demand for freedom. The success of the Prussian armies provided him with an opportunity to test this assumption. Once Austria had been subdued, he cajoled and bullied the secondary governments above the Main into joining Berlin in the establishment of the North German Confederation. It was the union of a giant and 21 pygmies, for the Hohenzollern kingdom constituted about four-fifths of the territory and population of the confederation. Executive authority was vested in a presidency held in accordance with hereditary right by the rulers of Prussia, who were to exercise the powers of their office with the assistance of a chancellor responsible only to them. The legislature was composed of a federal council, or Bundesrat, whose members were appointed by the state governments, and a lower house, or Reichstag, elected by equal manhood suffrage. Since Berlin had 17 votes out of 43 in the Bundesrat, it could easily control the proceedings with the support of a few of its satellites among the small principalities. Although the Reichstag could theoretically exercise considerable influence over legislation by granting or withholding its approval, parliamentary authority and party initiative were weak and untested, and Bismarck had little difficulty in piecing together a workable majority for his program by a strategy of divide and rule. His support came largely from a combination of moderate liberals and moderate conservatives willing to sacrifice theory for expediency. The federal constitution provided no bill of rights, no ministerial responsibility, and no civilian supervision over military affairs. But it introduced uniformity in currency, weights, measures, commercial practices, industrial laws, and financial regulations. In short, it created the economic unity that the middle class had long demanded and that helped reconcile it to the defeat of its hopes for greater political freedom.

The North German Confederation

Franco-German conflict and the new German Reich. The Seven Weeks' War, by creating a powerful new state in the heart of central Europe, abruptly altered the system of international relations on the Continent. Every government now had to reexamine its diplomatic and military position in the light of the establishment of the North German Confederation. No nation, however, was affected by the victory of the Hohenzollern armies as directly as France. Emperor Napoleon III had encouraged the outbreak of hostilities between Austria and Prussia on the assumption that both combatants would emerge from the struggle exhausted and that the Second Empire of France could then expand eastward against little resistance. The outcome of the war revealed how shortsighted such calculations had been. Instead of profiting from the conflict between Francis Joseph and William I, Paris suddenly confronted a strong and united German state that presented a serious threat to French interests. The imperial government was bound to regard this turn of events with suspicion and hostility. It sought to mitigate its discomfiture by seeking compensation in the Rhineland, Luxembourg, or Belgium. But Berlin succeeded in frustrating these plans, and the conviction began to grow in France that sooner or later a struggle with Germany would be unavoidable. The prospect of a new armed conflict was not unwelcome to Bismarck. He wanted to see national unification consummated by the entry of the southern states into the North German Confederation. Yet public opinion below the Main remained distrustful. Only a common patriotic struggle against foreign aggression might overcome the reluctance of the south to unite politically with the north. Thus in Berlin as well as in Paris there were reasons for seeking a test of strength. The immediate occasion came in the spring of 1870 with the candidacy of Prince Leopold, a relative of William I, for the throne of Spain. The ensuing controversy was cleverly exploited by Bismarck to provoke the French into initiating hostilities which both sides had wanted but for which the French government received the blame.

When France learned of Leopold's acceptance of the offer of the Spanish crown, there were wild protests in Paris and an immediate demand that Leopold be ordered to withdraw. On July 12 Leopold's father renounced the Spanish candidature on his behalf. This was not enough for the French government; it insisted that William I, as head of the Hohenzollern family, should promise that the candidature would never be renewed. This demand was presented to William at Ems by the French ambassador. Though William refused to give a promise, he dismissed the ambassador in a friendly enough way. But when the "Ems telegram," a report of what was happening, reached Bismarck, he shortened it for publication in such a way as to imply that William had refused to see the French ambassador again. The French used this as an excuse to declare war on Prussia on July 19.

Bismarck's calculation that a struggle waged ostensibly against the expansionist designs of Napoleon III would overcome particularism below the Main proved correct. The southern states joined the north in the Franco-German War, and the sense of unity engendered by the brotherhood of arms was soon enhanced by the intoxication of victory. The German troops won one battle after another in hard fighting along the frontier, until on September 2 they forced a large French army, headed by the emperor himself, to surrender at Sedan. The result was the establishment of a republican government in France, which continued to wage the struggle in the name of the old revolutionary ideals of 1793. The generalship of Moltke, however, was too much for the fierce determination of the new regime. Paris capitulated on Jan. 28, 1871, after a long and bitter siege, and on May 10 the Treaty of Frankfurt brought the war officially to a close. The Third Republic had to cede Alsace-Lorraine, pay an indemnity of five billion francs, and accept an army of occupation. It was a Carthaginian peace designed to crush a dangerous rival. The work of national unification in Germany, in the meantime, was successfully completed even before hostilities had ended. Bismarck had entered into negotiations with the southern states soon after the outbreak of war, determined to use patriotic fervour as an instrument for achieving political consolidation. The enthusiasm aroused in central Europe by the victory over France proved too much for the defenders of particularism. On January 18, while Prussian guns bombarded Paris, William I was proclaimed emperor of a united nation at military headquarters in Versailles. The governments below the Main joined the North German Confederation to form a powerful new Reich under the Hohenzollerns. Within a single lifetime, central Europe had completed the transition from cosmopolitanism to nationalism, from serfdom to industrialization, from division to union, from weakness to preponderance, from the Holy Roman Empire to the German Empire. (T.S.H.)

Bismarck's unification of Germany

Germany from 1871 to 1945

THE GERMAN EMPIRE, 1871–1918

The German Empire was founded on Jan. 18, 1871, in the aftermath of three successful wars by the north German state of Prussia. Within a seven-year period Denmark, the Habsburg monarchy, and France were vanquished in short, decisive conflicts. The empire was forged not as the result of an outpouring of nationalist feeling from the masses but from traditional cabinet diplomacy and agreement by the leaders of the states in the North German Confederation with the hereditary rulers of Bavaria, Baden, Hesse, and Württemberg. Prussia, occupying more than three-fifths of the area of Germany and having approximately three-fifths of the population, remained the dominant force in the empire until its demise at the end of another war in 1918.

At its birth Germany occupied an area of 208,825 square miles and had a population of more than 41 million, which

was to grow to 67 million by 1914. The religious makeup was 63 percent Protestant, 36 percent Roman Catholic, and one percent Jewish. The nation was ethnically homogeneous apart from a modest-sized Polish minority and smaller Danish and French populations. Approximately 67 percent lived in villages and the remainder in towns and cities. Literacy was close to universal because of compulsory education laws dating to the 1820s and '30s.

Domestic concerns. From its origins in 1871, the empire was governed under the constitution designed four years earlier by Otto von Bismarck, the Prussian prime minister, for the North German Confederation. This constitution reflected the predominantly rural nature of Germany in 1867 and the authoritarian proclivities of Bismarck, who was a member of the Junker landowning elite. There were two houses: the Reichstag, to represent the people, and the Bundesrat, to represent the 25 states. The former was composed of 397 members elected by universal manhood suffrage and a secret ballot. The constituencies established in 1867 and 1871 were never altered to reflect population shifts, giving rural areas a vastly disproportionate share of power as urbanization progressed. In theory the Reichstag's ability to reject any bill seemed to make it an important reservoir of power; in practice, however, the power of the lower house was circumscribed by the government's reliance on indirect taxes and the Reichstag's limited right to approve the military budget once every seven years. All legislative proposals were submitted to the Bundesrat first and to the Reichstag only if they were approved by the upper house. In addition, the legislative bodies were rarely consulted about foreign policy. Imperial ministers were chosen by and were responsible to the emperor rather than to the legislature. In comparison with the lower houses of the French and British parliamentary systems of the period, the Reichstag was a relatively impotent body.

The Constitution of 1867

A problem that was to plague the empire throughout its existence was the disparity between the Prussian and imperial political systems. In Prussia the lower house was elected under a restricted three-class suffrage system, a propertied franchise that allowed 15 percent of the male population to choose approximately 85 percent of the delegates. A conservative majority was always assured in Prussia, whereas the imperial franchise resulted in increasing majorities for the political centre and left-wing parties. William I (1797–1888) was both German emperor and king of Prussia. Apart from two brief instances the imperial chancellor was simultaneously prime minister of Prussia. Thus, the executives had to seek majorities from two separate legislatures elected by radically different franchises. A further problem was that government ministers were generally selected from the civil service or the military. They often had little experience with parliamentary government or foreign affairs.

The constitution had been designed by Bismarck to give the chancellor and monarch primary decision-making power. Universal suffrage had been proposed because of Bismarck's belief that the rural population would vote for either the Conservative or Free Conservative parties. The Progressives, a left-wing liberal party, were expected to do poorly in the two-thirds of Germany that was rural. Bismarck had not counted on new parties such as the Centre Party, a Roman Catholic confessional party, or the Social Democrats (SPD), both of which began participating in imperial and Prussian elections in the early 1870s. The Centre generally received 20–25 percent of the total vote in all elections. The SPD grew from 2 seats in the first imperial election to 35 by 1890, in which year the SPD actually gained a plurality of votes. Bismarck termed the Centre and SPD along with the Progressives *Reichsfeinde,* or enemies of the empire, because they sought each in its own way to change the fundamental conservative political character of the empire.

Beginning in 1871 he launched the Kulturkampf, a campaign in concert with German liberals against political Catholicism. Bismarck's aim was clearly to destroy the Centre Party. Liberals saw the Roman Catholic church as politically reactionary and feared the appeal of a clerical party to the one-third of Germans who professed Roman Catholicism. Both doubted the loyalty of the Catholic population to the Prussian-centred and, therefore, primarily Protestant nation. In Prussia, the minister of ecclesiastical affairs and education, Adalbert Falk, introduced a series of bills establishing civil marriage, limiting the movement of the clergy, and dissolving religious orders. All church appointments were to be approved by the state. As a result hundreds of parishes and several bishoprics were left without incumbents. Clerical civil servants were purged from the Prussian administration.

The Kulturkampf

The Kulturkampf failed to achieve its goals and, if anything, convinced the Roman Catholic minority that their fear of persecution was real and that a confessional party to represent their interests was essential. By the late



The German Empire, 1871–1918.

1870s Bismarck abandoned the battle as a failure. He now launched a campaign against the SPD in concert with the two conservative parties and many National Liberals. Fearing the potential of the Social Democrats in a rapidly industrializing Germany, Bismarck found a majority to outlaw the party from 1878 to 1890, although it could not constitutionally be forbidden to participate in elections. Party offices and newspapers were closed down and meetings prohibited. Many socialists fled to Switzerland and sought to keep the party alive in exile. During the 1880s Bismarck also sought to win the workers away from socialism by introducing legislation granting them modest pensions, accident insurance, and a national system of medical coverage. Like the Kulturkampf, the campaign against the SPD was a failure, and when the 1890 elections showed enormous gains for the *Reichsfeinde,* Bismarck began to consider having the German princes reconvene, as in 1867, to draw up a new constitution. The new emperor, William II (1859–1941), saw no reason to begin his reign with a potential bloodbath and asked for the 74-year-old chancellor's resignation. Thus, Bismarck, the architect of German unity, left the scene in a humiliating fashion, believing that his creation was fatally flawed. Indeed, his policy of supporting rapid social and economic modernization while seeking to avoid any reform of the authoritarian political system did lead to an atmosphere of persistent crisis.

The economy 1870–90. The empire was founded toward the end of two decades of rapid economic expansion, which saw the German states surpass France in steel production and railway building. By 1914 Germany was an industrial giant second only to the United States. After the establishment of the North German Confederation (1867) the impediments to economic growth were quickly removed. The usury laws and fetters on internal migration disappeared. A uniform currency based on gold was adopted by Bismarck and his National Liberal allies. An imperial state bank was created, and the tough regulations hindering the incorporation of joint stock ventures fell by the wayside. Combined with the euphoria over unification, these changes led to an unprecedented boom between 1870 and 1873. The *Gründerjahre,* as the years after unification were called, saw 857 new companies founded with a capital of 1.4 billion talers—or more new companies and investment in the private sector than in the previous 20 years. Dividends reached an astounding 12.4 percent. The railway system almost doubled in size between 1865 and 1875. Tens of thousands of Germans invested in stock for the first time to demonstrate both their patriotism and their faith in the future of the new German Empire.

The *Gründerjahre*

These halcyon years came to an abrupt end with the onset of a worldwide depression in 1873. The prices for agricultural and industrial goods fell precipitously; for six successive years the net national product declined. A sharp decline in profits and investment opportunities persisted until the mid-1890s. About 20 percent of the recently founded joint stock companies went bankrupt.

In agriculture, the deeply indebted Junker elite now faced severe competition as surplus American and Russian grain flooded the German market. Among the more immediate consequences of the crash was a burst of emigration from the overpopulated provinces of rural Prussia. During the 1870s some 600,000 people departed for North and South America; this number more than doubled in the 1880s. As a result of the depression social and economic questions increasingly preoccupied the Reichstag, while constitutional and political issues were put on the back burner.

It would be incorrect to draw the conclusion that the economy remained in the doldrums for an entire generation. While the 1870s and early 1890s were depressed periods, the 1880s saw significant recovery in industry, if not in agriculture. The British, who paid scant attention to Germany's emergence as an industrial power, began to respect their competitor during this decade.

In adjusting to the "Great Depression," Germany's leaders chose to return to a regulated or administered economy, after a generation of increasingly free trade. The hallmark of the new age was concentration; Germany became the land of big industry, big agriculture, big banks, and big government. The two areas in which the trend toward a controlled economy was most evident were tariff policy and cartelization. Cartel agreements, which were sanctioned by the state, apportioned markets, set standards for manufactured goods, and fixed prices. It is not coincidental that Germany, where the guild system prevailed into the 19th century, should have given birth to the cartel. Cartels arose rapidly in the steel, coal, glass, cement, potash, and chemical industries. Between 1882 and 1895 the total number of business enterprises grew by 4.6 percent, but the number employing more than 50 workers grew by 90 percent.

Cartelization

In 1878–79 Bismarck initiated a significant change in German economic policy, which coincided with his new alliance with the two conservative parties at the expense of the National Liberals. Tariffs were introduced on iron as well as on the major agricultural grains; the latter were raised in 1885 and again in 1887. This departure from liberal economic policy was necessitated by complaints from industrialists, estate owners, and peasants about the terrible impact the depression was having on their respective incomes. Only Britain held out against the protectionist tide that swept Europe in the 1880s. Bismarck's shift, nevertheless, had serious political implications. It signified his opposition to any further evolution in the direction of political democracy. The grain tariffs provided the Junker estate owners of Prussia, who constituted the main opposition to full political emancipation, with subventions that insulated them somewhat from the international market. Thus, the landed elite, major industrialists, the military, and the higher civil service formed an alliance to forestall the rise of social democracy, prevent further political liberalization, and make sure that the uncertainties of the market did not weaken the elites.

Foreign policy 1870–90. Until his resignation in 1890 Bismarck had a relatively free hand in the conduct of foreign policy. After three successful wars he saw his task as promoting peace and gaining time so that a powerful German Empire in the middle of Europe would come to be accepted as natural rather than as an interloper. The Prussian victories had led to great insecurity among the continental powers, who now adopted universal military service in imitation of Germany and maneuvered for defensive alliances so they would not find themselves isolated in the event of war. Bismarck's two areas of concern were the Balkans, where the disintegration of the Turkish Empire could easily lead to conflict between the Habsburg Monarchy and Russia, and France, which nurtured a desire for revenge against the German victors. Each might spark a general European conflagration that would inevitably involve Germany.

Although Bismarck is considered the foremost practitioner of realpolitik, his policies reflected a preference for dealing with monarchies and a disdain for parliamentary governments. In 1873 he negotiated the Three Emperors' League with Russia and Austria-Hungary. The league collapsed in the mid-1870s when rebellion broke out in Turkey's Slavic provinces. In 1877 Russia declared war on Turkey, leading both Britain and Austria-Hungary to express serious concern about Russia's expansionist war aims. When Russia imposed an annexationist peace on Turkey in the Treaty of San Stefano, Bismarck called for an international conference to reconsider the peace treaty and to forestall another military conflict. At the Congress of Berlin in 1878 Bismarck played the role of honest broker among the powers. Russia reluctantly accepted more modest gains and tensions dissipated.

But a conflagration had barely been avoided. Soon after the conference Bismarck negotiated an alliance with the Habsburg Monarchy (1879), which remained in effect through World War I. Thus he tied the fate of the youthful German Empire to the aged multinational empire that faced continuous problems from its many ethnic minorities. The chancellor had clearly opted for the Dual Monarchy over Russia should a war break out. The alliance gave him leverage in Vienna, and he steadfastly used it to prevent a war over the Balkans. His choice of the Habsburg Monarchy resulted from his fear that its dissolution would

The Dual Alliance

lead to Russian hegemony over the Polish, Czech, and other Slavic provinces of the empire. In addition, seven million Austro-German Catholics might seek admission to the German Empire, leading to a strengthening of the hated Centre Party.

Having a solid ally, Bismarck demonstrated his virtuosity by negotiating a revived Three Emperors' League in 1881. He now had influence in St. Petersburg as well as in Vienna to prevent a conflict over the Balkans. In 1882 Italy, fearing French hostility, joined the Dual Alliance with Austria-Hungary, making it into the Triple Alliance. On the surface Bismarck had triumphed. France had no allies for a war of revenge, and the alliance with the Habsburgs and Russia at once gave him influence with the two major adversaries in the Balkans.

The transient nature of this artistry soon became apparent. A crisis in Bulgaria inflamed Russo-Austrian relations, leading to the breakup of the revived league. Once again a war was avoided with Bismarck's intervention, but the Habsburgs and the Romanovs could no longer patch up their differences. Bismarck negotiated a separate Reinsurance Treaty with the tsar in 1887. Nevertheless, France and Russia began courting each other before Bismarck left office.

Between 1870 and 1890 Bismarck earned the respect of European leaders for his pacific policies. Apart from a few colonial additions in the mid-1880s, Germany under his guidance acted as a satiated power. The question remained whether this burgeoning industrial power led by the Junker and industrial elites would continue this policy while the other Western powers carved out world empires.

Politics 1890–1914. The political structure established by Bismarck in 1867 remained with scant change until the empire's demise in 1918. Leo von Caprivi, Bismarck's successor, was a political neophyte, having spent his entire career in the military. Given the disjuncture between the Prussian and German political systems (see above), Caprivi, surprisingly, sought to work with the parties of the centre and left, Bismarck's *Reichsfeinde.* With their support he reduced the grain tariffs and negotiated long-term trade treaties with Russia, the Dual Monarchy, and Romania. Food prices fell as a result, and industry flourished. National wealth rose rapidly, as did the standard of living of the industrial labour force. The Junker elite were outraged at Caprivi's willingness to sacrifice their interests in behalf of industry and labour. Utilizing their political power in Prussia and their access to the emperor, they were able to force his resignation in 1894, making his chancellorship the shortest before the war. After his resignation, the former general wrote to a friend, "In regard to the 'Junker agrarians' I see only evil, and it appears to me that a revolution by the agrarians is not impossible and for the moment more dangerous than a Social Democratic revolution."

Leo von Caprivi

Succeeding chancellors learned from Caprivi's fall that opposition to the landed elite was fraught with peril. Bernhard von Bülow, chancellor from 1900 to 1909, abandoned Caprivi's trade policy and resurrected the alliance of the agrarian and industrial elites.

To many observers the empire was edging toward a monumental crisis as it entered the 20th century. While its authoritarian political system was marked by paralysis, its economy was the most dynamic in Europe. With each election the increasingly urban electorate returned Social Democrats in growing numbers. By 1890 the Social Democrats (who had adopted a Marxist program of revolution at their Erfurt congress in 1891) received more votes than any other party, although four other parties won more seats. By 1912 they had more voters supporting them than the two next-largest parties combined. Both the Centre and Social Democrats were able to create parties with a mass base in German society. The Conservatives, National Liberals, and Progressives were more traditional parties, led by notables who were ill at ease in the world of populist politics. All three were declining, especially the Conservatives, who, despite flirting with anti-Semitism after 1893 by becoming a Christian party, fell to less than 15 percent of the vote by 1912. Many contemporary observers thought that a major crisis was impending between the recalcitrant elites and the increasing number of Germans who desired political emancipation similar to that of Britain and France.

While the Liberals and Conservatives declined in the Reichstag, a new phenomenon on the political scene was the rise of single-issue extraparliamentary interest groups. For the most part organizations such as the Pan-German League, the Navy League, the Farmers League, and the Colonial League were authoritarian in their politics and aggressively expansionist in foreign policy. Their constituencies were overwhelmingly middle-class and educated (except for the Farmers League), and they sought to influence the decision-making process both directly, by impressing the ministers with their strength, and indirectly, by supporting parties that adhered to their goals. Given the wealth and high status of their membership (professors were highly visible as leaders), they were unusually effective in achieving their goals. One of the striking characteristics of the empire was the support it received from the educated strata of the population, despite its elitist constitution.

The rise of interest groups

In the last election during the empire (1912) the Social Democrats scored a great victory, capturing 34.8 percent of the vote and 110 seats. On the local level they had begun to cooperate with the Progressives and occasionally with the Centre Party. Southern states such as Württemberg were moving toward full parliamentary government, and Alsace-Lorraine was given a surprising degree of self-government. Thus, there were some indications that the empire was evolving into a representative democracy. On the other hand, the states of Saxony and Hamburg adopted even more restrictive franchises than Prussia in the years before World War I. Above all Prussia and its Junker, military, and bureaucratic elites, supported by much of the professoriat, stood firm in opposition to any further democratization. Some historians have viewed the outbreak of the war in 1914 as an attempt by these elites to shore up their sagging position by a successful war and annexations, as Bismarck had done in the 1860s when the authoritarian Prussian state was besieged by a liberal opposition.

Economy 1890–1914. The speed of Germany's advance to industrial maturity after 1890 was breathtaking. The years from 1895 to 1907 witnessed a doubling of the number of workers engaged in machine building, from slightly more than one-half million to well over a million. An immediate consequence of expanding industrial employment was a sharp drop in emigration; from an average of 130,000 people a year in the 1880s, the outflow dropped to 20,000 a year in the mid-1890s. The surplus population continued to leave Prussia's eastern provinces, but the destination was the growing and multiplying factories of the Ruhr rather than the Americas. Earlier British fears of German competition now came to fruition. Only one-third of German exports in 1873 had been finished goods; the percentage rose to 63 percent by 1913. Germany came to dominate all the major continental markets except France.

The focus of national wealth as well as population shifted to the urban industrial sector by 1900. Only 40 percent of Germans lived in villages by 1910, a drop from 67 percent at the birth of the empire. Cities of more than 100,000 inhabitants accounted for one-fifth of the population in 1914, compared to one-twentieth at the time of unification. The application of intensive agricultural techniques led to a doubling in the value of all farm products despite a sharp decline in the rural population. Industry accounted for 60 percent of the gross national product in 1913.

The German working class grew rapidly in the Wilhelmian era. Total union membership reached 3.7 million in 1912, of which 2.5 million were affiliated with the socialist unions. Bismarck's welfare legislation covered some 13.2 million workers by 1911. Although German employers were extremely authoritarian and hostile to collective bargaining, the labour force did make significant economic gains. Between 1867 and 1913 the number of hours worked per year declined by 14 percent. Nearly every study of real income shows a rapid rise until 1902 and then a modest increase yearly thereafter. National income

per capita rose from 352 marks to 728 during the life of the empire. Despite these advances industrial workers lacked full political rights, which led a large number of them, including many Roman Catholic workers, to vote for the revolutionary socialist party.

While industrialization was rapid, it occurred only in certain sectors of the economy; other areas were only marginally affected. Some two million Germans persisted in traditional artisanal enterprises as the nation became an industrial colossus. While Germany was characterized by large Junker estates and cartels, it was also the nation of dwarf-sized farms (60 percent of farmers owned less than five acres) and the small workshop. German factories were larger and more modern than their British and French counterparts, but the preindustrial sector was more backward. During depressions those in the traditional trades often turned to anti-Semitism as an ideology that was at once patriotic and anticapitalist.

Foreign policy 1890–1914. Bismarck's successors rapidly abandoned the premises of his foreign policy. The Reinsurance Treaty of 1887 with Russia was dropped, leaving the empire more firmly tied to the Dual Monarchy and Russia free to conclude an alliance with France in 1894. Within four years Friederich von Holstein, a councillor in the political division of the foreign office, had weakened Germany's influence in the Balkans and allowed France to end its isolation. German overtures to Britain remained nugatory.

In 1895 the brilliant young sociologist Max Weber gave an inaugural lecture in Freiburg in which he pointed out that while Germany was establishing a nation-state belatedly the other powers had been founding world empires in Africa and Asia. Weber admonished his audience that Germany had to follow suit or become another Switzerland. Weber's contemporaries needed little convincing. From 1897 to 1912 William II, with his politically astute naval adviser, Alfred von Tirpitz, launched a policy of *Weltpolitik*. Germany's naval power went from being negligible to being second only to Britain's in little more than a decade. Nor did Germany build a navy simply to defend its coastline; rather, the new battleships were capable of challenging the other naval powers on the oceans. Tirpitz was a master of publicity, able to win much of the commercial and industrial middle class to his vision of a mighty empire, whose shipping lanes would be guarded by his fleet. The fleet was also expected to make the monarchy more popular and stem the growth of the left-wing parties. *Weltpolitik* proved to be a colossal failure. Germany came on the imperial scene late, when the choicest territories had already been occupied. Togo, part of New Guinea, and southwest Africa hardly seemed to justify the enormous expenditures on the navy. Moreover, Tirpitz's plans alienated Britain. Germany already had the most powerful army in the world when it fastened on becoming a great naval power. The British found this threatening and negotiated an alliance with Japan in 1902 and another one with France in 1904. In 1907 Britain settled its differences with Russia, and the Triple Entente (including France) was established. Germany now found itself surrounded by three major powers.

Weltpolitik

The more established powers found Germany meddling everywhere. This was particularly striking, because it followed two decades of Bismarck's policy of avoiding conflict. The Japanese objected to Germany's involvement in China in the 1890s. Russia watched as German power and influence were established in Turkey, its hereditary enemy. The French, of course, still harboured dreams of undoing their defeat in 1870. With Britain also alienated, Bismarck's nightmare of a coalition against the young upstart power, Germany, had become a reality. Bismarck had cre ated a system designed for his extraordinary talents; neither his successors nor the unstable young emperor, William II, shared the master's gifts. Parliament was more or less ignored, although some insightful critiques of *Weltpolitik* were to be heard in the Reichstag. Twice Germany's rulers sought to break up the alliance that was forming against the empire. In 1905 and 1911 Germany created crises over French penetration of Morocco. In each case the British and Russians stood firm, and, even though Germany gained concessions, the Triple Entente remained solid.

In 1912 the chancellor, Theobold Bethmann Hollweg, and his ministers reassessed the decade and a half of *Weltpolitik* and judged it a failure. The army, which had not been expanded since 1894, once again became the pride of the regime. Eastern Europe and the Balkans were now considered the most likely areas for German economic and political penetration. Since Italy had become unreliable, Austria-Hungary was the only ally to be counted on in the event of war. Any threat to the stability of the Dual Monarchy could leave Germany totally isolated.

The assassination of the Austrian heir to the throne, Archduke Francis Ferdinand, by Serbian terrorists in June 1914 augured poorly for the future of Austria-Hungary unless it showed resolve in dealing with the provocation. William II and Bethmann Hollweg urged strong measures against Serbia and reasserted their unconditional loyalty if war should eventuate. With Russia industrializing and the Dual Monarchy increasingly threatened by the aspirations of its minorities, time appeared to be on the side of the Triple Entente. Thus, if war was inevitable, the sooner it came the better. A localized conflict between Austria-Hungary and Serbia with a quick victory for the former would be desirable. If Russia chose to intervene to help its Slavic brother, Serbia, then a general European conflict would ensue. This was acceptable to the German government both because of its pessimism about the long-term strength of the Central Powers (*i.e.*, the German Empire and Austria-Hungary) and because the civilian population could be expected to rally to the war effort if tsarist Russia appeared to bear much of the responsibility. War, which inevitably gives the state increased power over civil society and demands sacrifices for the existing regime, may also have been attractive to elites fearing the growing strength of Germany's socialist movement.

World War I. During the first days of World War I Germans experienced a sense of bonding that had eluded them since the founding of the empire. Differences of class, religion, and politics were transcended as Germans flocked to their city centres to show their enthusiastic support for the impending conflict. Overwhelmingly, the parties, including the Social Democrats, voted for war credits. The euphoria of the early days masked Germany's dangerous situation. The Triple Entente commanded the seas, outnumbered Germany in population by three to one, and had access to the world's natural resources through their empires and close contact with the United States. Germany was immediately blockaded and had to rely on its own resources and those of contiguous nonbelligerents like The Netherlands, Denmark, and Switzerland. The Central Powers did have interior lines of transit, which was valuable in a two-front war. They also had a unified command structure—in contrast to the Triple Entente, where rivalries resulted in three different wars being fought simultaneously with little coordination.

General Alfred von Schlieffen, chief of the German general staff, had recognized Germany's vulnerability in a two-front war and saw the best hope in an overwhelming attack against France through Belgium. If all went according to plan, France's industrial region would be occupied in six to eight weeks and Paris itself surrounded. The slow-moving Russians would occupy Prussia's eastern rural provinces, facing only a modest-sized German military force. After France's concession of defeat following the occupation of its capital, whole armies would be shifted to the eastern front to drive the Russians out. Schlieffen died in 1913, and the plan was put in motion by General Helmuth von Moltke. As often happens in history, the plans of men may go awry in ironic ways. The western armies of Germany did, indeed, move through neutral Belgium but were stopped at the Battle of the Marne (September 1914) in northern France. Meanwhile, Paul von Hindenburg was reactivated at age 67 and sent with Major General Erich Ludendorff to halt the Russian advance into East Prussia. There the Germans unexpectedly defeated two large Russian armies at the Battle of Tannenberg (August 1914).

The Schlieffen Plan

The western front turned into a war of attrition as the two sides built opposing trenches from the Swiss border to the English Channel. For three and a half years neither

side moved more than 30 miles despite titanic battles at Verdun, the Somme, and Ypres. In the east the outmanned German forces, with the help of the Austrians, inflicted a series of costly military defeats on Russia, but, given the vast plains of eastern Europe, the Central Powers were unable to knock Russia out of the war until after the seizure of power by Lenin in the October Revolution of 1917. In 1916 Ludendorff and Hindenburg became joint heads of all German land forces and recognized, as had their predecessor Erich von Falkenhayn, that the war would be won or lost on the western front. With Italy (1915) and Romania (1916) entering the war on the side of the Triple Entente, the Central Powers faced an almost impossible situation in a war of attrition. The two generals became de-facto rulers of Germany and sought the mobilization of the whole society for total war. More than 11 million men, some 18 percent of the population, were in uniform, of whom almost two million were ultimately killed. Germany was unable to feed itself, and after the severe winter of 1916–17 malnutrition and even starvation were not uncommon.

On the diplomatic front the elites ruling Germany planned for vast annexations of Russian, Belgian, and French territory as well as for an African empire. The war costs were to be paid by the defeated powers of the Triple Entente. At no time during the war did the German government engage in serious negotiations to restore the sovereignty of Belgium or to return to the status quo ante. Nor were the allies very interested in a negotiated peace, but their situation was not as desperate as Germany's.

Ludendorff and Hindenburg adopted an all-or-nothing policy in regard to victory. They created an independent state of Poland in 1916, which prevented serious negotiations with Russia for a separate peace. They adopted submarine warfare in 1917, despite the knowledge that it would bring the United States into the war, because it offered a slim hope of quick victory if Triple Entente ships carrying men and supplies could be prevented from reaching France. Ludendorff also mounted a major offensive in April 1918, ignoring Woodrow Wilson's proposal of Fourteen Points for a future peace and failing to offer any peace terms of his own. When asked what would happen if the offensive failed, he replied, "Then, Germany will be destroyed."

In 1917 the Reichstag, following the lead of the Centre Party, passed a peace resolution based on no annexations. Social Democrats and Progressives rallied to support the resolution. The military and civilian leadership ignored the resolution and enforced a draconian peace on Russia and Romania in 1917–18. When the major battle in the west was brewing in April 1918, there were more than a million soldiers in the east to enforce the Treaty of Brest Litovsk with Russia.

Clearly, the military, agrarian, and industrial elites who ruled Germany viewed themselves as involved in two wars simultaneously, one against the Triple Entente and the other against the aspirations of the German people for full political emancipation. The latter conflict dictated victory at all costs on the military front. Defeat or a compromise peace on the battlefield would inevitably lead to democratization because of the sacrifices endured by many millions of workers, farmers, and artisans. In November 1914 Alfred Hugenberg, the major industrialist and subsequent ally of Hitler, told German entrepreneurs, "The consequences of the war will be unfavourable to employers and industry in many ways. One will probably have to count on a very increased sense of power on the part of the workers and labour unions, which will find expression in increased demands on the employer for legislation. It would, therefore, be well advised in order to avoid internal difficulties to distract the attention of the people and to give fantasies concerning the extension of German territory room to play." William II felt compelled to promise an eventual end to the restrictive Prussian franchise in his Easter message of 1917. Shortly thereafter the Fatherland Party was established with enormous support from the elites. Its program included a Carthaginian peace and maintenance of the Prusso-German political system.

The Ludendorff offensive of April 1918 made great breakthroughs in the west. But the effects of four years of attrition were apparent. The military did not have the reserves to take advantage of the initial gains. With almost a million fresh American troops in France, the Allies launched a counterattack that quickly gave them the initiative. Slowly the German forces began retreating. On August 8 the German army suffered a severe defeat in northern France, and not long thereafter William II installed a new, more liberal, government in Berlin, headed by Prince Max von Baden. The new ministers were informed that the war was virtually lost, and they were advised to seek an immediate armistice. Before the negotiations were successful, revolution broke out in the German navy and spread to the military and urban workers. William II was forced to abdicate as emperor of Germany and king of Prussia. The Social Democrats took power at this appalling moment of defeat, while the former military and civilian leaders sought to escape responsibility for the calamity. A civilian, Matthias Erzberger of the Centre Party, signed the armistice, which took effect Nov. 11, 1918. (For further discussion of World War I, see the article WORLD WARS, THE.) (Ke.B.)

THE GERMAN REPUBLIC, 1918–33

The republic proclaimed early in the afternoon of Saturday, Nov. 9, 1918, is often called the "accidental republic." When Friedrich Ebert, the leader of the so-called Majority Socialists, accepted the imperial chancellorship from Maximilian, Prince of Baden, it was with the understanding that he would do his utmost to save the imperial system from revolution. Ebert believed that the only way to accomplish this would be by transforming Germany into a constitutional monarchy. Elections would have to be held for a constituent assembly, whose task it would be to draw up a new constitution.

Defeat of revolutionaries, 1918–19. Ebert, however, was faced with a precarious situation. The dangers confronting him were mounting all over the country. Four and a half years of war and sacrifice were giving way to a war-weariness that was leaving the imperial system, as well as its emperor, discredited. Shortages of food and fuel had rendered the population vulnerable to the influenza epidemic sweeping Europe. On October 18 alone Berlin authorities had reported 1,700 influenza deaths. Independent Socialists in Munich had forced the abdication on November 8 of Bavaria's King Ludwig III and proclaimed a Bavarian socialist republic. The port cities along the North Sea and the Baltic Sea were falling into the hands of sailors' and workers' and soldiers' councils (*Räte*) in the wake of the naval mutiny at Kiel in early November. Berlin's radical leaders, Karl Liebknecht and Rosa Luxemburg, were eager to transform Germany into a republic of workers' and soldiers' councils (a *Räterepublik*) in imitation of the soviet republic being established by the Bolshevik leaders in Russia. As Ebert was accepting the reins of government in the Reichstag building on November 9, Liebknecht was addressing a rally of his own followers in front of the deserted Royal Palace about a mile away. Among Marxist revolutionaries the view was widespread that the Bolshevik revolution was merely the spark that would set off the worldwide proletarian revolution predicted by Karl Marx. Inevitably, that revolution would have to spread to Germany. Given this ideologically charged scenario, Liebknecht confidently anticipated his destiny to become the German Lenin.

While the Liebknecht rally was proceeding in front of the Royal Palace, an angry crowd was gathering before the governmental headquarters in the Reichstag building. Because Ebert had just left the building, his friend and fellow Majority Socialist Philipp Scheidemann felt called upon to address the crowd. To meet its inevitable demands for change and to forestall whatever Liebknecht might be telling his followers, Scheidemann in his speech used the phrase "Long live the German Republic!" Once said, the proclamation of a republic could not be withdrawn. Ebert was furious when he learned of Scheidemann's "accidental" proclamation, but he realized that there was no turning back. He spent the afternoon seeking partners to form a provisional government to run the newly pro-

claimed republic. By nightfall he managed to persuade the Independent Socialists, a party that in 1917 had split from the Majority Socialists over the continuation of the war, to provide three members of a provisional government. To gain their cooperation, Ebert had to agree to naming the provisional government the Council of Peoples' Commissars and to transforming Germany into a vaguely defined social republic. This promise notwithstanding, Ebert still hoped that elections to a constituent assembly would lead to the creation of a moderate democratic republic. The Independent Socialists, however, though not as radical as Liebknecht, held to their vision of a socialist *Räterepublik*. They hoped that workers and soldiers would elect a multitude of councils across the entire nation during the next weeks, which they assumed would establish the foundation for a genuinely socialist republic.

For the time being, however, Majority and Independent Socialists were joined in providing provisional governance for the defeated German nation, which everywhere seemed on the verge of collapse. Although the armistice of November 11 ended the fighting, it did not end the Allied blockade. The winter of 1918–19 brought no relief in the shortages of food and fuel, and the flu epidemic showed no signs of abatement. Soldiers returning from the military fronts by the hundreds of thousands were left stranded, jobless, hungry, and bitter—grist for the mill of revolution.

The push for revolution, led by an enthusiastic Liebknecht and a more reluctant Luxemburg, came on Jan. 6, 1919, encouraged by Soviet Russia and further prompted by fear that Ebert's plans for the election of a constituent assembly, scheduled for January 19, might stabilize the German situation. The Spartacists, now officially the Communist Party of Germany, initiated massive demonstrations in Berlin and quickly seized key government and communications centres.

The Spartacist revolt

The events of "Spartacist Week," as the radical attempt at revolution came to be known, demonstrated that Germany was not nearly as ripe for revolution as its leaders had thought it to be. As Luxemburg had feared, mass support among German workers for communism did not exist; most of them remained loyal to the Independent Socialists or to Ebert's more moderate and democratic vision of socialism. The German Army, moreover, had recovered its nerve and was determined to prevent a further move to the left. In December the army had begun secretly to train volunteer units drawn from the sea of soldiers returning from the front. These so-called Freikorps ("free corps") units became the basis for the dozens of small right-wing armies that during the next years roamed the country, looking for revolutionary activity to suppress. The Spartacist revolt, which was confined largely to Berlin, was put down within a week by some 3,000 Freikorps men. When Liebknecht and Luxemburg were captured on January 15, they were both shot at the initiative of Freikorps officers. Although sporadic revolutionary activity continued elsewhere in Germany during the next months, its failure in Berlin clearly marked its doom. The proclamation on April 4, 1919, of a *Räterepublik* in Bavaria revived radical fortunes only briefly; Freikorps units put down the radical Bavarian republic by the end of the month.

The collapse of the Spartacist revolt greatly enhanced the chances for Ebert's vision of Germany's future to prevail. Moreover, the meeting of a national congress of workers' and soldiers' councils in mid-December 1918, upon which the Independent Socialists had pinned their own hopes for creating a socialist republic, proved to be far less radical than expected; it did nothing to interfere with Ebert's plans to elect an assembly to draw up a democratic constitution. The elections on Jan. 19, 1919, which for the first time included women, produced a resounding victory for Ebert's conception of democracy. Three of every four voters gave their support to political parties that favoured turning Germany into a democracy. After months of turmoil Germany was to become a democratic republic. The assembly began its deliberations on Feb. 6, 1919, choosing to meet in Weimar, where it believed itself less vulnerable to radical political interference than in Berlin.

On Jan. 18, 1919, representatives of the powers victorious over Germany began the deliberations in Paris that would establish a European peace settlement. Germany's new democratic leaders placed high hopes in the prospects for this settlement. Woodrow Wilson's Fourteen Points seemed to promise Germans national self-determination as well as to encourage the efforts to transform Germany into a democracy. When the German constituent assembly met in Weimar for the first time on February 6, it immediately declared itself sovereign over all Germany. It selected a provisional government—with Ebert as president and Scheidemann as chancellor—whose first major task was to prepare for the expected invitation to Paris to negotiate a treaty of peace with the empire's former enemies.

But the invitation for a German delegation to come to Paris did not arrive until early April. Rather than being treated as a fellow—if fledgling—democracy, Germans soon learned that they were still viewed as the "bad boys" of Europe. Wilson's idealism had been forced to yield to still-fresh wartime resentments being articulated by the leaders of the French, British, and Italian delegations. There were to be no negotiations about a peace treaty. Germany would simply be handed a treaty and told to sign.

The Treaty of Versailles. The treaty in its final form contained many provisions that the Germans had fully expected. That Alsace-Lorraine was to be handed back to France was no surprise; nor were the small territorial adjustments along the border with Belgium. The plebiscite allowing the Danish population of northern Schleswig to choose between joining Denmark or remaining with Germany was unarguably consistent with the principle of national self-determination. But this principle, the Germans expected, would also justify a union between Germany and the Germans of what now remained of Austria after the collapse of the previous November. More serious to Germany was the stipulation that its coal-rich Saar region was to be taken over by the League of Nations and the coal given to France to aid its postwar reconstruction. Eventually a plebiscite was to allow Saarlanders to choose whether or not they wished to be returned to Germany.

On its eastern frontier Germany was forced to cede to the newly restored Poland the province of West Prussia, thereby granting Poland access to the Baltic Sea while losing land access to the province of East Prussia. Danzig was declared a Free City under the permanent governance of the League of Nations. Much of the province of Posen, which, like West Prussia, had been acquired by Prussia in the late 18th-century partitions of Poland, was likewise granted to the restored Polish state. Also lost to Germany as the result of a later plebiscite was a significant portion of coal-rich Upper Silesia.

Overseas, Germany was forced to yield control of its colonies. Although these colonies had proven to be economic liabilities, they had also been symbols of the world-power status Germany had gained in the 1880s and '90s. More immediately serious were the treaty's commercial clauses that took from Germany most of its foreign financial holdings and reduced its merchant carrier fleet to roughly one-tenth of its prewar size.

The treaty's provisions for disarming Germany were to be, the Allied leaders promised, merely the first step in a worldwide process of disarmament. To insure that Germany would not revive as a military power, its army was to be reduced to 100,000 men and would not be allowed to produce tanks, poison gas, or military planes. Moreover, Germany's frontier with France was to be permanently demilitarized; German military forces were to remain behind a line 31 miles east of the Rhine. The treaty also called for the dissolution of the German General Staff, the military command structure of the German Army, which the Allies believed to be the engine of German aggression. The navy, too, was to be dismantled and limited to 15,000 men, a half-dozen battleships, and 30 smaller ships. A prohibition on the building of submarines was absolute. Germany's compliance with the treaty's terms was to be assured by an Allied occupation of the Rhineland and the presence of the Inter-Allied Commissions of Control.

The most resented terms of the treaty, however, were the

so-called honour clauses: Articles 227 through 230 gave the Allies the right to try individual Germans, including the former emperor, as war criminals; Article 231, often called the war-guilt clause, provided the justification for Article 232, setting up a commission to collect reparation payments, the total of which was eventually set at 132 billion goldmarks (about 32 billion gold dollars). German bitterness over these honour clauses was nearly universal. Almost no German believed his country responsible for the outbreak of war in 1914. Technically, Article 231 did not declare Germany to be solely responsible for causing the war, merely "for causing all the loss and damage" suffered by the Allies in the war "imposed upon them by the aggression of Germany and her allies." Germans read it as an accusation of guilt, however, and understood it as the cynical product of victors' justice.

The German provisional government in Weimar was thrown into upheaval when it learned of the treaty's full terms. "What hand would not wither that binds itself and us in these fetters?" Scheidemann asked, and he resigned rather than accept the treaty. Army chief Paul von Hindenburg did the same, after declaring the army unable to resume the war under any circumstances. Only an ultimatum from the Allies finally brought a German delegation to Paris to sign the treaty on June 28, 1919, five years to the day since the assassination of Archduke Francis Ferdinand.

The Weimar Constitution. During the month following the signing of the treaty, the Weimar constituent assembly completed its drafting of a constitution for the new republic. The result was widely hailed as the most modern democratic constitution of its day. The document provided for a popularly elected president who was given considerable power over foreign policy and the armed forces. In Article 48 he was also given emergency decree powers to protect the republic from crises initiated by its opponents from either the left or the right. The president was empowered to nominate the chancellor, whose government required the confidence of the lower house of the parliament, the Reichstag, elected by universal suffrage and a system of proportional representation. An upper house, the Reichsrat, was composed of delegates appointed by the governments of the federal states, the *Länder.*

The most modern features of the Weimar constitution, the provisions for popular initiative and referendum, were designed to enable the electorate on its own initiative, by way of petition, to introduce bills into the Reichstag and to force the body to a vote on the proposal. If the bill was voted down, a national referendum was to be held to allow the electorate to pass the bill into law against the wishes of the Reichstag. Never again could a government afford to ignore the wishes of the voters.

The Weimar constitution was promulgated formally on Aug. 11, 1919, ending the provisional status of government in Germany that had begun with Scheidemann's proclamation of a republic the previous November. In September the government, judging the situation sufficiently safe in Berlin, returned to the capital. But it did not yet consider it sufficiently safe to risk the nationwide elections for president or for a Reichstag to replace the constituent assembly. Instead the assembly prolonged Ebert's provisional term as president for three years; elections for the Reichstag were delayed until June 1920.

Years of crisis, 1920–23. In its early years the new German democracy was beset by continuing turmoil. The Treaty of Versailles, quickly labeled "the *Diktat*" by the German public, galvanized the resentment that had accumulated during the war, much of which was turned back on the republic itself. Its enemies began to put the blame for the hated treaty upon the republic's socialist and democratic progenitors, whom they accused of having undermined Germany's efforts in the final stages of the war. A revived and radicalized right wing asked whether the German Army might not have been stabbed in the back by traitors on the home front. Racist circles took seriously the notorious *Protocols of the Learned Elders of Zion,* a fraudulent document fabricated in Russia in 1895 and published in Germany in 1920. It allowed the conclusion that all of recent history, including World War I, was the product of a conspiracy of Jews seeking to control the world. Roving Freikorps units contributed to the brutalization of German political life. In March 1920 one of these units, under the command of the former naval captain Hermann Ehrhardt, succeeded in briefly establishing its control of the government in Berlin. This so-called Kapp Putsch, named after the conservative politician Wolfgang Kapp, who had planned it, was put down not by the army but by a general strike of Berlin's socialist and communist workers. A similar right-wing putsch in Bavaria was successful, however, and from that point forward radical groups of the right found protection and a degree of nurture in this south German state. By the end of 1922 there had been nearly 400 political assassinations, the vast majority of them traceable to the right wing. The victims included prominent politicians such as Matthias Erzberger, a signatory to the Armistice of 1918, and Walther Rathenau, the foreign minister.

Kapp Putsch

The difficulties in which the new democracy found itself were already reflected in the June 1920 elections to the first Reichstag. The Weimar coalition parties, the SPD (socialist), the Centre Party (Roman Catholic), and the Democrats, which in January 1919 had received over 75 percent of the vote, this time managed to win only 43.5 percent. Contributing to the problems Weimar faced in the early 1920s was the escalating rate of inflation that was eventually to destroy the German mark. Although the inflation was rooted in the huge debt that Germany had amassed in financing the world war, the hyperinflation of 1923 was triggered by the French-Belgian military occupation in January 1923 of the German industrial district in the Ruhr valley. The occupation occurred in retaliation for Germany's having fallen behind in its reparation payments and was intended to force German industry to provide compensation for the French and Belgian losses. Rather than accede quietly to the humiliation of occupation, the German government urged workers and employers to close down the factories. Idle workers were paid for the next months with a currency inflating so rapidly that printers gave up trying to print numbers on bills. By mid-1923 the German mark was losing value by the minute. A loaf of bread that cost 20,000 marks in the morning would cost 5,000,000 marks by nightfall. Restaurant prices went up while customers were eating. Workers were paid twice a day. On November 15, when the collapse came, it took 4.2 trillion German marks to buy a single American dollar.

The hyperinflation of 1923

The social and political cost of the hyperinflation was high. Scholars note that the inflation did more to undermine the middle classes than the presumably socialist revolution of 1918. A lifetime of savings would no longer buy a subway ticket. Pensions planned for a lifetime were wiped out completely. Politically, the hyperinflation fueled radicalism on both the left and the right. The Communists, badly set back by their failure in January 1919, believed the prospects for a successful revolution to be greatly improved. In Munich the leader of the small National Socialist German Workers' Party, Adolf Hitler, used the turmoil to fashion an alliance with other right-wing groups and attempt a coup in November 1923 that sought to use Bavaria as a base for a nationalist march on Berlin. His hope was to overthrow the democratic system of Weimar that he believed was responsible for Germany's political and economic humiliation. Neither the radicals of the right nor those of the left succeeded in imposing their will. In the short run they did not succeed because of ineptitude and miscalculation; in the long run they failed because the government sponsored a currency reform that led to a restabilization of the mark and also decided to end its policy of passive resistance in the Ruhr in exchange for an end to the occupation and a rescheduling of the reparation payments it owed to the Allies.

The Weimar Renaissance. Amid the political and economic turmoil of the early 1920s Germany's cultural and intellectual life was flowering. The so-called Weimar Renaissance brought to Germany the fulfillment of the Modernist revolution, which in the late 19th century had begun to transform the European aesthetic sensibility. The Modernist rejection of tradition seemed to suit perfectly the need of many Germans for new meanings and values

to replace those destroyed by the war. "A world has been destroyed; we must seek a radical solution," said the young architect Walter Gropius upon his return from the front in late 1918. In 1919 Gropius became the founder and first director of the Bauhaus school of design in Weimar, the most important institution in Germany for the expression of Modernism's aesthetic and cultural vision. Bauhaus artists rejected the styles of the previous century in their entirety. Along with other artists of the 1920s, they believed themselves to be creating a new world through their painting, poetry, musical compositions, theatrical productions, and architectural constructions. The legacy of German Modernism in general, and of the Bauhaus in particular, is most immediately evident in the stark steel-and-glass high-rise buildings whose clear and clean lines have come to dominate the skylines of the world's cities. Moreover, the paintings and the sculptures decorating them, as well as the design of the furniture and the lighting fixtures, are heavily influenced by the aesthetic principles articulated in Weimar during the 1920s.

Beyond the Bauhaus, painters like George Grosz, Max Beckmann, and Otto Dix pursued an artistic objective described as Expressionism; *i.e.,* they were interested in depicting their emotional responses to reality rather than reality itself. In music the rejection of tonality by composers such as Arnold Schoenberg, Anton von Webern, and Alban Berg broke a centuries-old tradition in music. At the juncture between popular and serious music the composer Kurt Weill collaborated with the poet Bertolt Brecht to create in 1928 *The Threepenny Opera,* a bitterly satiric musical play in which the world of modern capitalism was equated with that of underworld gangsterism. In films such as *The Cabinet of Dr. Caligari,* distorted sets and unusual camera angles probed for disturbing truths behind the surface appearances of reality.

Not everyone was pleased by the Modernist attack on tradition. Opera performances and theatrical productions were often interrupted by angry audiences. Siegfried Wagner, the son of the composer Richard Wagner, deplored a Modernist version of his father's *The Flying Dutchman,* calling the production an example of "cultural bolshevism." Other artists, the novelist Thomas Mann, for example, chose to remain above the fray in order to examine their world from the Olympian heights of German *Kultur,* an effort for which Mann in 1929 won the Nobel Prize for Literature.

Years of economic and political stabilization. The financial recovery that began with the restabilization of the German currency in late 1923 was undergirded in 1924 when the Allies agreed to end their occupation of the Ruhr and to grant the German government a more realistic payment schedule on reparations. A committee of the Allied Reparations Commission headed by the American financier and soon-to-be vice president Charles Dawes had recommended these changes as well as urged the Allies to grant sizable loans to Germany to assist its economic recovery. The Dawes Plan marked a significant step in the upswing of the German economy that lasted until the onset of the Great Depression. The 800 million gold marks in foreign loans had by 1927 enabled German industrial production to regain its 1913 prewar high. That same year the Reichstag contributed to the vital need for social and class reconciliation by voting for a compulsory unemployment insurance plan. Reconciliation on the political level seemed achieved in 1925 when the 77-year-old Hindenburg was elected to succeed the deceased Friedrich Ebert as president. Although no democrat, the aged field marshal took seriously his duty to support the constitution and the republic.

The Dawes Plan

The guiding spirit in German foreign policy from 1924 through 1929 was the foreign minister, Gustav Stresemann, who firmly believed that Germany was more likely to gain relief from the harshness of Versailles by trying to fulfill its terms than by stubbornly continuing to resist them. Stresemann's efforts ushered in what came to be known as "the era of fulfillment." It began in December 1925 when Germany signed the Locarno treaties, in which it guaranteed to maintain the new postwar boundaries with France and Belgium and to submit to international arbitration any boundary disputes that might arise in the east with Poland or Czechoslovakia. Germany was formally readmitted into the family of nations by being granted membership in the League of Nations in September 1926. In 1928 Germany became party to the most dramatic symbolic gesture of postwar reconciliation, the Kellogg-Briand Pact, which promised to outlaw the resort to aggressive war; this agreement was signed by nearly all the world's major countries during the next year.

The economic and political stabilization of the German republic seemed to be reflected in the May 1928 elections to the Reichstag. The antirepublican parties of the left and right together received only 13 percent of the total vote, with the Communists receiving 10.6 percent and the Nazis taking only 2.6 percent. Germany's reintegration into the international political structure was reflected also in the decision in early 1929 by the Allied Reparations Commission to make a final disposition of the reparations question. The committee appointed to make recommendations in this matter was headed by Owen D. Young, an American business executive. The Young Committee proposed that German reparations be reduced to about 37 billion gold marks, less than one-third of the 1921 total, and that payments be stretched until 1988. It also called for the dissolution of the Reparations Commission and for an immediate end to what remained of the Allied occupation of the Rhineland.

The German government, seeing the obvious advantages in the Young Plan, officially accepted its terms in August 1929. Right-wing parties of the opposition, however, saw the plan as nothing less than a renewal of Germany's humiliation. Led by the German National Peoples' Party (DNVP) and its leader Alfred Hugenberg, the press and movie-industry lord, the nationalist opposition seized upon the constitutional processes for popular initiative and referendum in order to force the government to reverse its acceptance of the plan. To run the opposition's anti-Young Plan campaign, Hugenberg engaged Adolf Hitler, the leader of the apparently moribund NSDAP (National Socialist German Workers' Party). The objective was to force the German government into a repudiation of the reparations debt as well as of the war-guilt clause of Versailles upon which the debt rested. German signatories to the Young Plan, moreover, were to become liable to the charge of treason. The right wing's initiative did force the Reichstag into reconsidering its approval of the Young Plan, but to no avail. The national plebiscite that necessarily followed found only 13.8 percent of the voters favouring the objectives of the right wing. The bitterness of the campaign, however, may have contributed to the illness and death during the campaign of Stresemann.

The end of the republic. An unintended effect of the anti-Young Plan campaign was to give widespread public exposure to the little-known Adolf Hitler. Hitler used his access to the Hugenberg-owned press empire and to its weekly movie newsreels to give himself and his Nazi movement national publicity. An additional assist to Hitler's career came on Oct. 29, 1929, with the stock market crash on Wall Street, an event that signaled the onset of what quickly became a worldwide depression. The crash had an immediate effect in Germany as American investors, anxious about their financial position, began withdrawing their loans to Germany. German indebtedness to these investors had by 1929 reached nearly 15 billion marks. Prices on the German stock exchanges fell drastically during the last month of the year. Business failures multiplied. Early in 1930 Germany's second-largest insurance firm collapsed. Unemployment rose to three million during the course of the year. By the winter of 1932 it reached six million. Germany's industry was working at no more than 50 percent of its capacity, and the volume of German foreign trade fell by two-thirds between 1929 and 1932.

The stock market crash of 1929

The first critically important political effect of the economic crisis came in March 1930 when the government coalition fell apart over the rising cost of maintaining the unemployment program adopted in 1927. The SPD, representing labour, and the Peoples' Party (VP), representing business, were unable to agree on the size of the government's contribution to the fund, and therewith the gov-

ernment fell. When a new coalition could not be found, parliamentary democracy in Germany came to an end.

Heinrich Brüning

President Hindenburg was forced by the situation to invoke his emergency powers (Article 48), which he used to appoint Heinrich Brüning of the Catholic Centre Party as chancellor. For the next two years, until May 30, 1932, Brüning governed without a parliamentary majority, deriving his authority from the powers residing in the office of President Hindenburg. However well-intentioned, Brüning's deflationary economic policies were unable to stem the tide either of the depression or of its social and political ravages. His fateful decision to call for Reichstag elections in September 1930, moreover, inadvertently opened the door for the enemies of Weimar democracy. Together the Nazis and the Communists gained nearly one of every three votes cast. The 18 percent acquired by the Nazis represented seven times the total number of votes they had received in 1928; the 13 percent won by the Communists represented a healthy gain over the 10 percent of 1928.

Although bitterly opposed to each other, the Nazis and Communists during the next two years succeeded in mobilizing the political and economic resentments generated by the depression. Hitler's charismatic appeal and the youthful energies of his movement were attractive to large segments of a populace fearful of being ruined by economic and social disaster. Hitler's record as a war veteran lent authority to the hypernationalism he expressed in racist terms. His identification of the Jew as the enemy responsible for all of Germany's ills, be it the defeat of 1918, the Treaty of Versailles, the reparations, the inflation, or now the depression, seemed plausible to many eager to find a scapegoat. The power of Hitler's appeal was reflected in the party's growing membership lists (1929: 170,000 members; 1932: 1,378,000 members) and in the swelling ranks of the party's paramilitary SA (Sturmabteilung), the infamous storm troopers.

The Communists, unlike Hitler, found it difficult to extend their support beyond the German working classes. Their political flexibility, moreover, was limited by the fact that they increasingly fell under the control of Stalin. Nonetheless, their self-confidence was greatly enhanced because the depression seemingly confirmed their analysis of the inevitable collapse of capitalism. Hitler and National Socialism they perceived merely as products of the last phase of capitalism.

The winter of 1931–32 witnessed the depression at its depths. Unemployment was still rising; the succession of business failures resembled rows of falling dominoes. Brüning, helpless in the face of these problems, was dubbed "the hunger chancellor" by his critics. Some hope of breaking through the political impasse was offered by the series of critical state legislative elections scheduled for the spring of 1932 and by the presidential election required at the expiration of Hindenburg's first term. The 84-year-old Hindenburg was with great difficulty prevailed upon to seek a second term. The year 1932 was to be one of continuous election campaigning.

Hitler chose to challenge Hindenburg for the presidency. Although Hindenburg was eventually reelected, a runoff was necessary, and Hitler was able to take 37 percent of the popular vote. His larger aim, however, had been to make himself the leading, or only, candidate for Brüning's position as chancellor. Hindenburg did choose to replace Brüning in May 1932, but rather than Hitler he named the political dilettante Franz von Papen. Desperate to find for himself a base in parliament, Papen called for Reichstag elections in July. The result was a disaster for Papen and another triumph for the Nazis, who took 37 percent of the vote, the largest total they were ever to acquire in a free election. The Communists won 15 percent of the vote. Thus the two parties dedicated to destroying German democracy held a majority in the Reichstag. Still, Hitler did not get the chancellorship. In November Papen called for another Reichstag election in the hope of gaining parliamentary backing. Again he failed, although the Nazi vote fell by 4 percent. The Communist vote, on the other hand, rose to nearly 17 percent. In early December, when Hindenburg decided to replace Papen, he again ignored Hitler, choosing instead a friend from the army, General Kurt von Schleicher.

In the Nazi camp there was bitter frustration at the end of 1932. The party was deeply in debt and demoralized by the year's endless campaigning. Putschist elements in the party, never persuaded that elections could bring the party to power, were growing increasingly restive. So deep was the frustration that on December 7 Gregor Strasser, second only to Hitler in the party, broke with the Nazis and retired from politics.

THE THIRD REICH, 1933–45

The Nazi revolution. When Hitler finally became chancellor on Jan. 30, 1933, it was not on the crest of a wave of popular support but as the result of backroom political intrigue by Schleicher, Papen, and the president's son, Oskar von Hindenburg. Only Hitler, they believed, could bring together a coalition with Hugenberg's DNVP and possibly the Centre Party that could command a majority in the Reichstag. They assured the reluctant president that Hitler's radical tendencies would be checked by the fact that Papen would hold the vice-chancellorship and that other conservatives would control the crucial ministries, such as those of war, foreign affairs, and economics. The Nazis themselves were restricted to holding the chancellorship and the powerless ministry of the interior. As a sop to the Nazis, Hermann Göring was granted ministerial status but given no portfolio; yet, significantly, he became interior minister in the state of Prussia, which gave him control over the largest police force in Germany.

The Nazis brought with them an ideology that purported to champion the common man, whom they portrayed as victimized in a world controlled by Jews. Anti-Semitism and notions of German racial superiority were at the core of this ideology, which, in its particulars, was also a catalog of resentments that had accumulated in German society since November 1918. Heading the list were the humiliations associated with Versailles, but not far behind were resentments of big business, big banks, big department stores, and big labour, as well as resentments of the divisiveness and inefficiencies seemingly being fostered by political parties.

Neither the 25-point party program (1920) nor Hitler's autobiographical *Mein Kampf* (1925) contained clear conceptions of how the German world would be structured under the Nazis, but Hitler and his propagandists had communicated clearly that the changes would be fundamental and come at the expense of Germany's racial enemies. Racially superior Germans were to be gathered into a tightly knit *Volksgemeinschaft,* or racial community, in which divisions of party and class would be transcended in a spirit of racial harmony, a harmony that would necessarily exclude people of inferior blood. The logic of this construct required a solution to what the Nazis called "the Jewish problem." At the very least it called for a reversal of the trend, more than a century old, of Jewish assimilation into the allegedly superior German nation and into German cultural and economic life. As for Germany's position in international affairs, Hitler had long spoken of Germany's need for living space (*Lebensraum*) in the east. First, however, there was the continued need to break the chains of the hated Treaty of Versailles.

Whether the Nazis would ever get a chance to implement their ideological objectives depended, when Hitler became chancellor, upon whether they would be able to tighten their initially tenuous hold on the reins of power. Liberals, socialists, and communists remained bitterly opposed to Hitler; important segments of business, the army, and the churches were to varying degrees suspicious of the measures he might take. It was a combination, finally, of Hitler's daring and brutality, of the weaknesses of his opponents, and of numerous instances of extraordinary good luck that allowed him to establish his totalitarian dictatorship. When the Centre Party refused to join the Nazi-DNVP coalition in January 1933, Hitler demanded elections for a new Reichstag. The elections of March 5, 1933, were preceded by a brutal and violent campaign in which Nazi storm troopers under the command of Ernst Röhm figured prominently. Hitler was also able to take

advantage of the Reichstag fire (probably the work of a lone and deranged Dutch communist) of February 27 to suspend civil liberties and arrest Communist as well as other opposition leaders. Despite this campaign of terror the Nazis did not win a majority, gaining only 43.9 percent of the total. The 8 percent acquired by the DNVP, however, was sufficient for the two parties to wield a majority in the Reichstag. At its first meeting on March 23 this Reichstag, under great pressure from the SA and Heinrich Himmler's SS (Schutzstaffel; "Protective Echelon"), voted in favour of an Enabling Act that gave Hitler power to ignore the constitution and to give his decrees the power of law.

The decree powers were the pseudolegal base from which Hitler carried out the first steps of the National Socialist revolution. Within two weeks of the passing of the Enabling Act, Nazi governors were sent out to bring the federal states into line, and a few months later the states themselves were abolished. On April 7, 1933, the civil service, including the universities, began to be purged of socialists, democrats, and Jews. On May 2 the trade unions were disbanded and replaced by what the Nazis called a Labour Front. In the meantime Göring had begun refashioning the political arm of the Prussian police into a secret political police (Gestapo [Geheime Staatspolizei]) to serve the Nazi cause, a process that was being duplicated by Himmler with the Bavarian police.

The brutality with which Hitler proved willing to meet any presumed challenge to his authority became dramatically evident in his ordering the June 30, 1934, murders of the SA leadership. Röhm's storm trooper street thugs had provided useful muscle during the party's long years of struggle, but their continuing penchant for unruliness, Hitler feared, could invite the army's intervention and therewith his own overthrow. To head off this possibility, Hitler engaged the loyal Himmler, who used his SS during the so-called Night of Long Knives to purge the SA of dozens of its top leaders, including Hitler's longtime friend Ernst Röhm. The penultimate step in Hitler's seizure of power came on Aug. 2, 1934, when, upon the death of President Hindenburg, he appropriated for himself the powers of the presidency and combined them with his own as chancellor. The final step came in February 1938 when Hitler took personal command over the three branches of the German armed forces.

The totalitarian state. The main purpose of the Nazi revolution, the goal to which power was to be put, was the establishment of the *Volksgemeinschaft.* Its creation required the purification and increase of the German race as well as its biological separation from the Jews, whose infusion of evil into the German bloodstream, the Nazis said, served to pollute and undermine Germany's well-being. Nazi efforts to purify the German race found a degree of scientific respectability from the new science of eugenics, the racial hygiene movement that flourished widely in the early decades of the 20th century. Nazi leaders spoke of their efforts to "reconstruct the German race." A Law for the Protection of Hereditary Health (July 14, 1933) allowed for the eventual sterilization of as many as two million people deemed unworthy of propagating. A Marriage Subsidy Law of July 1933 was calculated to stimulate the birth rate by granting loans to newly married couples; these loans would be forgiven incrementally with the birth of each additional child. The Nazi idealization of mothers and the celebration of motherhood as a special service to the Reich had the same objective. Hitler spoke of an eventual doubling of the German population through these measures. The most notorious of the steps taken to purify the German race was also a milestone in the anti-Jewish legislation promulgated by the Nazis: the infamous Nürnberg Laws of September 1935, which forbade marriage or sexual relations between Jews and Germans and assigned to Jews a lower class of citizenship.

Nazi efforts to solve the "Jewish problem" were ultimately products of a vicious anti-Semitism that propelled the Nazi system toward increasingly extreme measures of persecution. SA terrorism, legislation expelling Jews from the civil service and universities, boycotts of Jewish businesses and professional people, and the eventual expropriation of Jewish-owned properties had by 1938 led to the emigration of roughly half of the 1933 Jewish population of 500,000 people. Until the massive display of violence against Jews and Jewish property known as Kristallnacht (Night of Broken Glass), during the night of Nov. 9–10, 1938, it still seemed possible for some Jews to remain in Germany, albeit in severely circumscribed circumstances. With Kristallnacht the Nazis ushered in a new level of persecution. In its aftermath Hitler put Hermann Göring in charge of Jewish policy. It became Göring's job to coordinate the numerous party and governmental agencies competing for control over—and profit from—the persecutions of the Jews.

Kristallnacht

To exclude Jews from the *Volksgemeinschaft* was as critical a need as attracting the German working classes to it, thereby undoing the long-standing alienation of the largely socialist-minded German worker from the nationalist consensus. Any success in this effort depended heavily upon the Nazis' ability to provide employment for the millions of jobless whom they inherited in 1933. Public spending on rearmament and on public works projects such as the superhighway, or autobahn, network helped create many of the jobs so desperately needed. By 1937 Germany was beginning to suffer from a labour shortage. It was equally important to inculcate workers with a sense of being an integral part of a racially based national community. For this the Nazis devised an elaborate program of subsidies for leisure-time activities for workers. Called "Strength through Joy," the program subsidized workers' vacations; it made possible excursions to mountain or seaside resorts and offered the possibility of cruises in the Mediterranean or the Baltic Sea. By providing leisure activities for the workers until then reserved for their economic and social superiors, the government attempted to integrate them into the *Volksgemeinshaft.* Workers were also given the opportunity to purchase an automobile, the Volkswagen, a symbol of wealth and status still largely reserved in the early 1930s for the upper classes. Hitler once said that Henry Ford had done more than anyone else to obliterate class differences in America.

Foreign policy. Hitler kept tight control over foreign affairs, formulating himself both the strategy and the tactics calculated to achieve his goals. The immediate objective was to reestablish Germany's position in world affairs; by this Hitler meant ending the humiliations attending the Treaty of Versailles, such as the demilitarized Rhineland and the limitations on German armaments. The chains of that treaty needed to fall with a loud clang, he said. The larger objective, the one he had spoken about since his entry into politics in the early 1920s, was the conquest for Germany of *Lebensraum.* Hitler believed that this space needed to be acquired in the east, at the expense of the Soviets, so as to secure for Germany the Ukrainian "breadbasket" and open up vast territories for German colonization. Hitler found justification for such conquests in his notions of German racial superiority over the Slavic peoples who inhabited the lands he coveted. Furthermore, he saw the Bolsheviks who now controlled Russia as the vanguard of the world Jewish conspiracy. Control of this territory was to become the foundation for Germany's economic and military domination of Europe and eventually, perhaps, of the world.

No such domination or expansion was possible without war, of course, and Hitler did not shrink from its implications. His rearming of Germany, begun in secret in 1933, was made public in March 1935 when he announced the creation of an air force and the reintroduction of general military conscription to provide the manpower for 36 new divisions in the army. In June of that same year he signed an agreement with the British that allowed a German naval buildup of up to 35 percent of Britain's surface naval strength and up to 45 percent of its tonnage in submarines. On March 7, 1936, he moved German forces into the demilitarized Rhineland. Versailles was dead. Neither the British nor the French had lifted a finger in its defense, choosing instead to sign agreements expressly negating the terms of Versailles. By 1936 Hitler was spending 10.2 billion marks on rearmament, and Göring was placed in charge of a so-called Four-Year Plan to prepare the Ger-

man economy for war. On Nov. 5, 1937, Hitler gathered his general staff and admonished them to be prepared for war in the east no later than 1942 or 1943.

As a cover for his true intentions during the first years of power, Hitler spoke long and often of his desire for peace. All he wanted, he said, was a Germany allowed its rightful place in world affairs. As evidence of his pacific intentions he signed in January 1934 a 10-year nonaggression pact with Poland. A truer picture of his intentions became evident in July 1934, however, when Hitler encouraged the Nazi party in Austria to attempt an overthrow of the government of Chancellor Engelbert Dollfuss. Dollfuss was shot and killed, but the coup attempt was badly managed. Benito Mussolini's movement of Italian troops to the Austrian border forced Hitler to back away from supporting his Austrian partners.

It was from behind a cloak of anti-Bolshevism, carefully maintained by Propaganda Minister Joseph Goebbels, that Hitler managed to delude France and Britain, as well as the United States, leery of being drawn into another European quarrel, into accepting his claim to be the West's last bulwark against Bolshevik expansion. His agreement with the Japanese in 1936, the Anti-Comintern Pact, was directed against the Third Communist International. A year later Mussolini, after a year of German-Italian cooperation in aiding General Francisco Franco's rebel forces in the Spanish Civil War, added his signature to the pact.

Hitler made two dramatic foreign policy moves in 1938 that helped clarify to the world the extent of his less-than-pacific intentions. In March he annexed Austria to the Reich, justifying the move as a fulfillment of the principle of German national self-determination. Britain and France stood by quietly at this additional repudiation of the Versailles treaty. Next, Hitler engineered a diplomatic crisis with Czechoslovakia, claiming Czech mistreatment of its German minority in the Sudetenland. Against considerable opposition from his own military, Hitler was determined to go to war with the Czechs. Only through the intervention of Britain's Prime Minister Neville Chamberlain, who offered to come to Germany to appease Hitler and who managed to persuade the Czechs to yield to all of Hitler's demands, was war avoided. Chamberlain's intervention resulted in a four-power conference at Munich in late September 1938, at which the Italians, the French, and the British ceremoniously handed the Sudetenland over to Hitler. In diplomatic parlance ever since, the word "Munich" has come to symbolize caving in to the demands of a dictator.

Annexation of Austria

Certain that Britain and France would do nothing to stop him, Hitler, after Munich, decided to move up his timetable for conquests in the east. On March 15, 1939, Hitler seized what remained of Czechoslovakia, reshaping its pieces into a Bohemian and Moravian protectorate and a nominally independent state of Slovakia. Within a week he annexed from Lithuania the city of Memel and the surrounding countryside, territory lost to Germany as a consequence of Versailles. When Britain and France countered these moves by issuing their guarantee to Poland, clearly next on Hitler's agenda, a furious Hitler ordered his military to prepare an invasion of Poland. Poland was critical to Hitler's long-range strategy for the conquest of *Lebensraum* in the east; any invasion of the Soviet Union required that Polish territory be available as a staging area. Until the British-French guarantee he had hoped to enlist Poland, mostly through bombast and threat, as an ally in an attack on the Soviet Union. After the defeat of the Soviet Union, he believed Poland could be dealt with summarily. When Poland refused to play the role he had assigned it, Hitler began looking for allies in his resolution of the Polish problem. In late May 1939 he signed his Pact of Steel with Mussolini. Then came the great diplomatic revolution of the 20th century. Shortly after signing the pact with Italy, Hitler put out feelers to Stalin about the possibility of dividing eastern Europe, Poland included, into German and Soviet spheres of influence. The negotiations led to the notorious German-Soviet Nonaggression Pact of Aug. 24, 1939, a pact that secretly sealed Poland's doom and delayed Hitler's unbending determination to conquer *Lebensraum* from the Soviets. Thus prepared, on Sept. 1, 1939, Hitler launched his invasion of Poland. Two days later Britain and France declared war on Germany.

World War II. World War II is appropriately called "Hitler's war." The first two years were so extraordinarily successful that Hitler came close to realizing his aim of establishing a German hegemony in Europe. But his victories did not prove to be the essential components of a strategic conception that secured victory in the long run. Nonetheless, the early successes were spectacular. After the defeat of Poland within a month, Hitler turned his attention westward. He believed that it was necessary to defeat Britain and France before he could again turn eastward to the territories that were to become the "living space" for his new empire. The attack on the western front began in the spring of 1940. Hitler took Denmark and Norway during the course of a few days in April, and on May 10 he attacked France, along with Luxembourg, Belgium, and The Netherlands. Once again his armies achieved lightning victories; Luxembourg, Belgium, and The Netherlands were overrun in a few days, and France capitulated on June 21. Only the British, now alone, obstructed Hitler's path to total victory in the west.

Britain, Hitler determined, could be taken out of the war with air power. German bombers began their attack in August 1940, but the British proved intractable. The vaunted German air force failed to bring Britain to its knees partly because of the strength of the British air force, partly because the German air force was ill-equipped for the task, and partly because the British had been able since February 1940 to read German code. Yet Hitler had been so confident of a quick victory that even before the attack began he had ordered his military planners to draw up plans for an invasion of the Soviet Union. The date he set for the invasion was May 15, 1941.

Although the defeat of the Soviet Union was central to Hitler's strategic objective, he twice during the early months of 1941 allowed himself to be sidetracked into conflicts that delayed his invasion. In both instances he felt obliged to rescue his ally Mussolini from military difficulties. Mussolini had invaded Greece in October 1940, despite the fact that he was already in difficulty in North Africa, where he was unable to cut off Britain's Mediterranean lifeline in Egypt. In February 1941 Hitler decided to reinforce Mussolini in North Africa by sending an armoured division under the command of General Erwin Rommel. When Mussolini's invasion of Greece also bogged down, Hitler again decided to send reinforcements. To reach Greece, German troops had to be sent through the Balkan countries, all of them officially neutral. Hitler managed to bully these countries into accepting the passage of German troops, but on March 27 a coup in Yugoslavia overthrew the government, and the new rulers reneged on the agreement. In retaliation Hitler launched what he called Operation Punishment against the Yugoslavs. Yugoslav resistance collapsed quickly, but the effect was to delay for another month the planned invasion of the Soviet Union.

Invasion of the Soviet Union

When the invasion of the Soviet Union finally came, on June 22, 1941, it did so with both campaigns against the British, across the English Channel and in the Mediterranean, still incomplete. Hitler was prepared to take the risk that fighting on multiple fronts entailed because he was convinced that the war against the Soviet Union would be over by the onset of the Russian winter. The spectacular German advances during the first weeks of the invasion seemed proof of Hitler's calculation. On July 3 his army chief of staff wrote in his war diary that the war had been won. The German Army Group North was approaching Leningrad; Army Group Centre had broken through the Soviet defenses and was rushing toward Moscow; and Army Group South had already captured vast reaches of the Ukraine. The prospect of capturing the summer harvest of the Ukraine along with the oil fields of the Caucasus led Hitler to transfer troops driving toward Moscow to reinforce those operating in the south.

Hitler's generals later considered this decision a turning point in the war. The effect was to delay until October the drive toward Moscow. By then an early winter had set in, greatly impeding the German advance and finally bringing

it to a halt at the outskirts of Moscow in early December. Then, on December 6, the Soviets, having had time to regroup, launched a massive counteroffensive to relieve their capital city. On the following day the Japanese, nominally Germany's ally, launched their attack on the U.S. naval base at Pearl Harbor in Hawaii. Although they had not bothered to inform Hitler of their intentions, he was jubilant when he heard the news. "Now it is impossible for us to lose the war," he told his aides. On December 11 he declared war on the United States.

Though his plans for a quick defeat of the Soviet Union had not been realized, Hitler's troops at the end of 1941 controlled virtually all of the European territory of the Soviet Union. They stood at the outskirts of Leningrad and Moscow and were in control of the entire Ukraine. To prepare for what would now have to be the campaign of 1942, Hitler dismissed a number of generals and assumed himself the strategic and operational command of the armies on the eastern front.

At the high point of Hitler's military successes in the Soviet Union, members of the Nazi leadership were, with Hitler's understanding, feverishly planning for the new order they intended to impose on the conquered territories. Its realization called both for the removal of obstacles to German settlements and for a solution to the "Jewish problem." Nazi planners were drafting an elaborate scheme, General Plan East, for the future reorganization of eastern Europe and the western Soviet Union which called for the elimination of 30 million or more Slavs and the settlement of their territories by German overlords who would control and eventually repopulate the area with Germans. During the fall of 1941 Heinrich Himmler's SS expanded and refurbished with gas chambers and crematoriums an old Austrian army barracks near the Polish rail junction at Auschwitz. Here was to continue, with greater efficiency, the mass murder of Jews that had begun with the June invasion, when SS murder squads began rounding up Jews and shooting them by the thousands. The assurance of victory in the east, the heartland of European Jewry, gave Nazis the confidence that the "Jewish problem" could be rendered amenable to a "final solution." Experts suggest that ultimately somewhere between five and six million Jews were murdered in the death factories of eastern Europe. At least an equal number of non-Jews died of murder and starvation in places like Auschwitz, including two and a half million Soviet prisoners of war and countless others from the eastern European nationalities.

The success of Nazi armies until the end of 1941 had made it possible to spare German civilians on the home front from the misery and sacrifices demanded of them during World War I. Hitler's imagination, however, was haunted by the memory of the collapse of the home front in 1918, and to avoid a repetition the Nazis undertook to loot the occupied territories of food and raw materials as well as labour. Food shortages in Germany were not serious until late in the war. Women were allowed to stay at home, and the energies of the German work force were not stretched to their limits because eventually some seven million foreign slave labourers were used to keep the war effort going.

Through much of 1942 an ultimate German victory still seemed possible. The renewed offensive in the Soviet Union in the spring at first continued the successes of the previous year. Once again Hitler chose to concentrate on the capture of the Caucasus and its oil at the expense of the Moscow front. The decision made inevitable a major battle over the industrial centre at Stalingrad (now Volgograd). Elsewhere, by midsummer of 1942, Rommel's Afrika Korps advanced to within 65 miles of Alexandria in Egypt. In the naval battle for control of the Atlantic sea lanes, German submarines were able to maintain their ability to intercept Allied shipping into mid-1943.

By early 1943, however, the tide had clearly begun to turn. The great winter battle at Stalingrad brought Hitler his first major defeat. His entire Sixth Army was killed or captured. In North Africa Rommel's long success ended in late 1942 when the British broke through at El Alamein. At the same time a joint British-American force landed in northwestern Africa on the coast of Morocco and Algeria. By May 1943 the German and Italian forces in North Africa were ready to surrender. That same summer the Allies broke the back of the German submarine campaign in the Atlantic. On July 10 the Allies landed in Sicily. Two weeks later Mussolini was overthrown, and in early September the Italians withdrew from the war.

The addition of an Italian front made the rollback of German forces on all fronts that much more likely. In the Soviet Union, German forces were stretched across 2,500 miles. They had lost their air superiority when Allied bombing raids on German cities forced the withdrawal of large numbers of fighter planes. Allied bombings reached a high point in midsummer when a raid on Hamburg killed 40,000 of its inhabitants. Other cities suffered similar raids. Shortages of food, clothing, and housing began to afflict German cities as inevitably as did the Allied bombers.

The rollback of German forces continued inexorably during 1944. On June 6 the Allies in the west launched their invasion of France across the English Channel. In the east the Soviet Army was advancing along the entire 2,500-mile front. By the end of the year it stood poised on the eastern frontiers of prewar Germany. In the west, British and American troops stood ready to attack across the western borders.

On the German home front, 1944 became a year of acute suffering. On July 20 an officers' plot, part of a long-simmering opposition to Hitler from within German military and civilian circles, was carried out, but Hitler managed to escape the dramatic attempt on his life practically unharmed. He attributed his survival to his having been selected by fate to succeed in his mission of restoring Germany to greatness.

Fate did not again intervene on Hitler's behalf. In mid-January of 1945 he withdrew underground into his bunker in Berlin where he remained until his suicide on April 30. By that time Soviet soldiers were streaming into Berlin. All that remained of the Reich was a narrow wedge of territory running southward from Berlin into Austria. (For further discussion of World War II, see WORLD WARS, THE.)

With the Soviet army in control of Berlin and the western Allies within striking distance to the west and the south, there was no prospect of dividing them. Nonetheless, when Hitler's successor, Grand Admiral Karl Dönitz, sought to open negotiations for a surrender a few days after Hitler's death, he still hoped that a separate surrender to the British and Americans in the west might allow the Reich to rescue something from the Soviets in the east. The western Allies, fearful of any move that might feed the suspicions of Stalin, refused to consider the German proposal, insisting that a German surrender be signed with all of the Allies at the same time. Early in the morning of May 7, 1945, a German delegation came to U.S. General Dwight D. Eisenhower's headquarters in Rheims, Fr., and at 2:41 AM signed the surrender documents. Despite the fact that a Soviet major general signed for the Soviet Union, Stalin insisted that a second surrender ceremony be arranged in Soviet-occupied Berlin. This second surrender was signed in a Berlin suburb the following afternoon. (K.A.Sc.)

The era of partition

As a legacy of unconditional surrender at the end of World War II, a truncated Germany was divided into four zones of Allied military occupation; Berlin, the capital between 1871 and 1945 of the unified German Reich, was similarly divided. In 1949 the zones of occupation of France, the United Kingdom, and the United States were merged to form the Federal Republic of Germany (commonly known during the era of partition as West Germany), with its capital at Bonn. The three western sectors of Berlin also had a separate administration, beginning in 1948, constituting *Land* Berlin (commonly called West Berlin). West Berlin's political affiliation was with West Germany, although it was a detached exclave of territory at a distance of 110 miles (180 kilometres). Also in 1949 the Soviet zone became the German Democratic Republic (Deutsche Demokratische Republik; commonly known as East Germany). The Democratic Republic regarded the former Soviet sector of Berlin as an integral part of its ter-

ritory, and indeed as its capital. Even after the formation of the two states, the four Allied powers retained certain rights, including the stationing of troops; these rights proved vital to the West in ensuring the survival of West Berlin during and after the 1948–49 Soviet blockade.

West German recovery

West Germany's recovery from total economic and political prostration and the devastation of its cities and capital industries at the end of World War II was of such dramatic proportions as to become a modern legend. The *Wirtschaftswunder* ("economic miracle") of the 1950s soon made it the world's fourth largest economy after the United States, the Soviet Union, and Japan and one of the world's largest trading nations. Its currency became among the strongest in the world.



Germany, 1952–90.

Domestically, the West German state grew into one of the most stable of Western democracies. The three major political parties largely overcame their origins along class, regional, or sectarian lines. In large measure because of the 5 percent of votes required for representation in the federal parliament (Bundestag) and provincial diets, extremist parties of the right and left did not succeed in sending delegates to Bonn after the election of 1957. The West Germans as citizens moved far from their accustomed apathy or even perverse boast of being apolitical to becoming intensely involved and well-informed on the issues, domestic and foreign, that affected their lives.

The Democratic Republic was established as a socialist state of the Stalinist type and became a member of the Council for Mutual Economic Assistance (Comecon) and of the Warsaw Pact military alliance. It had the most advanced economy and the highest living standards of all the Soviet-bloc states, but industrially it lagged behind the Western level. The five states (*Länder*) established in East Germany by the victorious Allies of 1945 were replaced in 1950 by 15 administrative districts (*Bezirke*), each centred on a major city, including East Berlin. These districts lacked powers of independent decision, being intended merely to apply the policies of the central administration. Unlike other Soviet-bloc countries, East Germany could make no exclusive appeal to national sentiment as the basis of loyalty and unity; instead it offered the proclaimed political, economic, and social superiority of the socialist state. To outside observers, the centralized and dictatorial regime of the Socialist Unity Party (a forced union of the Social Democratic Party with the Communists), buttressed by the omnipresent State Security Police (Stasi), appeared unassailable. But the precarious nature of this foundation was demonstrated by its swift and unforeseen collapse under popular pressure in late 1989, most strikingly symbolized by the opening of the Berlin Wall on November 9. At free elections on March 18, 1990, the majority of East Germans, by voting for the counterpart of the West German Christian Democratic Union, signaled their wish to join the Federal Republic at the earliest possible moment (see below, *Germany reunited*). (T.H.El./G.H.K./Ed.)

THE FEDERAL REPUBLIC OF GERMANY

In a declaration of June 5, 1945, the Allies proclaimed that all central government authority in Germany had ceased to exist and that governmental powers had been taken over by the military governments of Britain, the United States, and the Soviet Union. Each was to administer its zone of occupation independently, with the proviso that all matters affecting Germany as a whole should be dealt with by the Allied Control Council. The Potsdam Conference of July–August 1945 corroborated the agreements that the Allied powers had reached during the war years; for the foreseeable future Germany was to be divided into three zones of occupation—later four, when France was accepted as an occupying power—run by military governments. Under the guidance and control of the military governments, representative institutions, beginning on the level of local governments, were gradually created, and the reestablishment of democratic parties was permitted. The Soviet Union was the first occupying power to begin the process of democratization, although with the obvious intention of securing a key position in the new political order for the German Communist Party (Deutsche Kommunistische Partei). The Western powers were somewhat slow to follow, but by the autumn of 1945 they were prepared to accelerate the process of democratization. The most important step was the creation of various German *Länder* (states), largely cutting across the traditional territorial divisions inside Germany (Prussia officially ceased to exist in February 1947). The new *Länder,* which with a few territorial alterations exist to the present day, initially provided the arena for the reemergence of democratic politics in Germany.

Allied zones of occupation

Although the occupying powers agreed in principle that the unity of Germany should be preserved, the Allied Control Council proved incapable of acting as a joint body to coordinate policies in the four occupation zones. Differences between the Western powers and the Soviet Union increased on almost all major issues, in particular on the issues of reparations and economic reconstruction. Neither did the Germans themselves succeed in bringing about a joint approach. A conference of all German *Länder,* including those from the Soviet occupation zone, on June 6, 1947, proved fruitless. The foreign ministers of the four powers meeting in London in the autumn of 1947 again failed to solve the deadlock. The final division of Germany was precipitated by the decisions of the Western powers in February 1948 to integrate the three Western occupation zones into a single economic unit, to develop it with the assistance of funds from the Marshall Plan, to subject the Ruhr district to international control, and to opt for a federal solution through the creation of a separate state in the Western zones. The currency reform implemented on June 20, 1948, by Ludwig Erhard as the head of the Bizonal Economic Administration, the only central governmental organization in the Western zones, amounted to a formal breach of the four-power agreements and provoked the Soviets to establish a blockade of West Berlin. This was a last, dramatic attempt to halt the setting up of a separate West German state. Subsequently the Allied powers encouraged the German political authorities and party leaders to set up a West German state, albeit on strictly federalist lines, and to work out a new constitution.

Creation of West Germany

Formation of the new state. The Basic Law (Grundgesetz) of the Federal Republic was passed by the Parliamentary Council on May 8, 1949, and four days later was approved by the Western Allied military governors with certain reservations, notably the exclusion of West Berlin, which had been proposed as the 12th *Land* of the federation. The 11 *Länder,* then, were Schleswig-Holstein, Hamburg, Bremen, Lower Saxony, North Rhine-

Westphalia, Rhineland-Palatinate, Hesse, Bavaria, Baden, Württemberg-Baden, and Württemberg-Hohenzollern (the last three were merged in 1952 to form Baden-Württemberg, and in 1957 Saarland became the 10th *Land*). When the Occupation Statute came into force in September 1949, the Allied military government was replaced by an Allied High Commission. In November the Petersberg Agreement made further concessions on dismantling industries and allowed the building of a merchant fleet. The Federal Republic became a direct member of the Ruhr Authority and the Council of Europe and was able to establish consular and commercial relations abroad.

The constitution drawn up for the Federal Republic was not unlike that of the Weimar period (1919–33), but it provided far greater democratic safeguards and established certain basic individual rights. The positions of president and chancellor were virtually reversed; in the new constitution the president was elected for a five-year term by a federal assembly made up of both legislative houses. The emergency powers available to the president under the Weimar constitution were now to be wielded by the chancellor and his cabinet. The chancellor, nominated by the president, had then to be elected by absolute majority in the Bundestag, or lower house. This house was elected by direct universal suffrage for a four-year term. The Bundesrat, or the upper house, represented the *Länder* governments: no legislation could be enacted without its consent.

First federal elections

The first federal election was held on Aug. 14, 1949. Led by Konrad Adenauer, who had been chairman of the Parliamentary Council, the Christian Democrats (a coalition of the Christian Democratic Union and its Bavarian counterpart, the Christian Social Union) won the most seats, followed closely by the Social Democrats. Of the other eight parties represented in the Bundestag, the most important were the Free Democratic Party, the conservative German Party, the federalist Bavarian Party, and the Communists. An alliance with the Free Democrats and the German Party enabled Adenauer to be elected chancellor and to form a coalition government. Theodor Heuss, of the Free Democratic Party, was elected president of the republic in September.

The chancellorship of Adenauer. *Foreign policy under Adenauer, 1949–63.* Adenauer, long a champion of European cooperation, assumed the role of foreign minister, in addition to his chancellorship, when the Foreign Office was established in 1951. Later, when the position fell to others, important decisions remained reserved to the chancellor. Adenauer's avowed aim was the reunification of Germany, but he regarded this as the Allies' responsibility, and continued differences between the Western and the communist powers made its achievement unlikely. Priority had to be given, in his view, to the firm integration of the Federal Republic into the West. Accordingly, Stalin's proposal in 1952 that Germany be reunited at the price of abandoning its ties to the West and with strict neutrality was refused, and the division of Germany into separate states became an established fact. Adenauer supported the Western alliance under U.S. leadership and worked to establish a federal relationship among the countries of western Europe. The Social Democrats sharply opposed this order of priorities until 1956 and also opposed the idea of German rearmament, pressed on the Federal Republic by the United States.

To prevent the reemergence of a German national army, Sir Winston Churchill (of Britain) and René Pleven (of France) proposed the setting up of a European Defense Community, in which German units would be integrated under European command. The Social Democrats consistently opposed this project until the elections of 1953 gave the government a two-thirds majority in both houses, which meant that, had they needed to amend the constitution on the rearmament issue, they could have done so. But the European Defense Community project was turned down by the French in August 1954, and, following agreements made in Paris in October 1954, the Federal Republic became, in May 1955, a sovereign state and a full member of the North Atlantic Treaty Organization (NATO), pledged to establish its own armed forces. Great Britain joined with the six member countries of the European Coal and Steel Community (established in 1951) in a Western European Union, which was both to supervise German rearmament and to ensure the maintenance of sufficient British forces on the continent.

Membership in NATO

Above all, Adenauer worked for closer European integration, and in 1955 West Germany was prominent in the discussions that led in 1957 to the signing of the Rome Treaties establishing the European Atomic Energy Community and the European Economic Community (EEC). The Social Democrats voted for their ratification.

Outside Europe, West Germany tried hard, especially among the emergent nations of Africa and Asia, to establish itself as a noncolonial power able to aid their economic development. Israel was a special case, receiving more than three billion deutsche marks in material reparation for Germany's treatment of the Jews under the Nazi government. Throughout the 1950s West Germany was on excellent terms with the United States, especially after the 1956 involvement of France and Britain in the Suez. A peak of cordiality was reached in 1963 when President John F. Kennedy visited West Germany and West Berlin.

Adenauer went to Moscow in 1955 and agreed to establish diplomatic relations with the Soviet Union, in return for the repatriation of nearly 10,000 German prisoners of war still held in the Soviet Union. A constant problem was the two countries' divergent views about the German Democratic Republic. Moscow considered that there were two German states, but the Federal Republic would not recognize the legitimacy of the East German government. Although a trade agreement was signed with the Soviet Union in 1958, a crisis developed over Berlin later in that year, when the Soviets demanded the withdrawal of Western troops and wanted to declare Berlin a free city. A series of international meetings failed to solve the conflict, and in 1961 the Soviets authorized the building of the Berlin Wall, separating the eastern from the western sectors.

Domestic affairs under Adenauer, 1949–63. The election of Adenauer meant the end of the Social Democrats' plans to establish a socialist system in West Germany. Instead, the system of a "social market economy" was firmly established, providing the base for a speedy economic recovery and, eventually, for what came to be known as the "German economic miracle." The most urgent internal problem for Adenauer's first administration was the resettlement of refugees. At the end of the war, some eight million Germans had been expelled from Polish territories or from the Sudetenland; there were others from the Balkans, and each year until 1961 tens of thousands of additional refugees arrived from East Germany. The presence of these refugees put a heavy social burden on the Federal Republic, but their assimilation proved surprisingly easy. Many of the refugees, indeed, were skilled, enterprising, and readily adaptable, and their labour proved an important factor in West Germany's economic recovery.

German refugees

In the elections of September 1953, the Christian Democrats won virtually half the seats in the Bundestag. Economic progress, Adenauer's personal prestige, and the desire to remain close to the protection of the West were the chief reasons for this success. The Social Democrats remained the largest minority. The Free Democrats, who as junior partners in Adenauer's coalition could neither attack its record nor claim much credit for its achievements, lost seats, and the German Party and the All-German Bloc, a specifically refugee party, each won a small number. Adenauer formed a second coalition government with the Free Democrats, the German Party, and the All-German Bloc. Heuss was reelected president in July 1954.

Conflicts soon developed. The trade unions and the Social Democrats launched a campaign against proposals for rearmament and conscription, although they were not able to prevent the passage of a law authorizing a volunteer army. The Conscription Bill was passed two years later, and the first recruits were called up in 1957.

During this term of office, Adenauer lost the effective support of some of the smaller parties in his coalition. Most of the All-German Bloc went into opposition in 1955, and the next year, after a bitter quarrel over an electoral law, the Free Democrats also left the government. But any further erosion of Adenauer's power was averted by

external events. The brutal suppression of the Hungarian rising by the Soviet Union in 1956 caused a swing back to the support of a government firmly in alliance with the West, while showing up as impractical various plans for reunification with East Germany that some of the Social Democrats and Free Democrats had been entertaining. The government was able to point to impressive social and economic achievements when launching its campaign for the elections of 1957. As a result, the Christian Democrats once again increased their proportion of seats in the Bundestag. Adenauer formed his third coalition in alliance only with the German Party.

Again, rearmament proved a bone of contention, the Social Democrats and the trade unions this time protesting against the possible equipment of the West German army with nuclear weapons. They also called for the withdrawal of foreign (that is, NATO) troops from West German soil. But the trade unions refused to call a strike over the issue; the constitution allowed no referendum on it; and in 1958 the federal armed forces received dual-purpose missiles capable of carrying atomic warheads. By the end of 1960, there were more than 290,000 men in the West German forces.

The year 1959 saw the first attempt to challenge Adenauer's power from within the Christian Democratic Union. The crisis was provoked when, as Heuss's second term of office was drawing to an end, Adenauer proposed to have Erhard nominated to succeed him as president, thereby eliminating him from any contest for the chancellorship. Erhard refused to agree, and, in a surprise move, the party induced Adenauer himself to accept nomination for the presidency. But, when he saw that the party would then make Erhard chancellor, Adenauer reversed his decision. Though, in the course of the crisis, Adenauer had cast grave aspersions on Erhard's political experience and reliability, he was able easily to crush the revolt in the party. The candidate finally elected president, in July 1959, was Heinrich Lübke, a Christian Democrat.

Adenauer's attempt to exclude Erhard

A notable change in the policy of the Social Democrats began in November 1959, with their adoption of the Bad Godesberg Program. This policy document showed them to have completely abandoned Marxism, while even socialization was scarcely mentioned. Roughly a year later, at Hannover, the party conference confirmed Willy Brandt, the mayor of West Berlin, as its candidate for the chancellorship. Brandt reversed the party's policy on the only major issue that now separated it from the government by accepting both conscription and the possibility that German troops might be armed with nuclear weapons.

The elections of 1961 returned only three parties to the Bundestag: the Christian Democrats with the greatest number of seats, the Social Democrats, and the Free Democrats. Having lost their absolute majority, the Christian Democrats allied themselves with the Free Democrats to form a government. Although that party's leader, Erich Mende, wanted a new chancellor, Adenauer was in fact reelected, after the Free Democrats had elicited an understanding that he would retire before the parliamentary term expired.

The chancellorship of Erhard (1963–66). Adenauer resigned in October 1963, and Erhard was elected chancellor, forming a coalition cabinet similar to Adenauer's but including Mende as vice-chancellor and minister for all-German affairs. Lübke was reelected president in July 1964.

The Federal Republic's relations with the West were complicated by the problem of reconciling "Atlanticism," European unity, and the Franco-German treaty—particularly as France under President de Gaulle became more estranged from NATO and insisted on the exclusion of Britain from the Common Market. Further difficulties arose over West Germany's status in any plan for a "multilateral nuclear force" within NATO; France wished to restrict control of its nuclear deterrent to itself, while Britain did not want to allow West Germany authority in decisions about the use of nuclear weapons. The plan, however, was abandoned in December 1965 when Erhard had talks in Washington, D.C., with President Lyndon B. Johnson. German-French relations remained at a stalemate, and relations with Britain, temporarily improved by a state visit from Queen Elizabeth II in May 1965, were clouded by disputes about liability for the financial upkeep of British troops stationed in West Germany.

Problems within NATO

Although the results of the election of September 1965, when the Christian Democrats won an unexpectedly large number of seats, were regarded as a personal victory for Erhard, his government ran into difficulties. For the first time since the currency reform of 1948, the growth rate of the West German economy slackened, and in 1967 industrial production actually fell. Widespread fear of economic recession, coupled with the lack of clear political leadership, led to two developments. In the country at large, the National Democratic Party—a neo-Nazi right-wing group—won some support, especially in Bavaria and Hesse, and in October 1966 the Free Democrats, worried about their own future as junior partners in a government in difficulty, withdrew from Erhard's cabinet.

The chancellorship of Kiesinger (1966–69). The withdrawal of the coalition partner left the cabinet without a parliamentary majority and the Christian Democrats without a leader who could command adequate parliamentary support. In November Kurt Georg Kiesinger, previously minister-president of Baden-Württemberg, became the candidate for the chancellorship and, after complicated negotiations, effected an unprecedented political alliance with the Social Democrats, called the "Great Coalition." Kiesinger was elected chancellor, and Brandt became vice-chancellor and foreign minister.

Kiesinger sought a détente with the East and an improvement of relations with France. Ambassadors were exchanged with the Soviet Union in 1967, but the Soviet occupation of Czechoslovakia in 1968 caused renewed tension. Friction with France not only continued but was increased by Brandt's warm endorsement of Britain's application to join the EEC, which was vetoed in December 1967 by de Gaulle. Fresh conflict arose over the currency crisis of November 1968, when the West Germans not only resisted pressure from France and other countries to revalue the deutsche mark but also recommended the devaluation of the French franc, a step de Gaulle refused to take.

Despite the fact that the Kiesinger government presided over a return of economic growth, it encountered vociferous opposition from both the right and the left. In June 1969 the problem of trials of "war criminals" had to be faced again, since the regulations extended in 1965 would soon run out. A time limit for the bringing of charges of genocide was abandoned, and the limit for murder was extended to 30 years from the end of the war.

Early in 1969, President Lübke announced his resignation, and the decision to hold the presidential election in West Berlin provoked reprisals from the East German authorities. The elections were held, nevertheless, and the new president was Gustav Heinemann, a Social Democrat.

The social-liberal coalition (1969–82). On Oct. 21, 1969, Willy Brandt became West Germany's first Social Democratic chancellor following elections that resulted in a significant shift to the political left. Brandt acted quickly to forge a coalition with the Free Democrats, who hitherto had been effectively left out of the government. A new economic policy attempted to give an equal say to employers and trade unions on the basis of what came to be called concerted action.

The most important achievement of the coalition under Brandt's leadership was *Ostpolitik* ("eastern policy"). The new government abandoned the so-called Hallstein doctrine, which forbade diplomatic relations with all countries entertaining diplomatic relations with East Germany. Instead, it sought to come to terms with its communist neighbours to the east. In 1970 West Germany signed a treaty with Poland ratifying existing borders, and in 1973 a similar treaty with Czechoslovakia was signed. In August 1970 Brandt signed a treaty with the Soviet Union confirming the existing frontiers and pledging the renunciation of force to achieve political ends. It was to be ratified by the Bundestag upon the successful conclusion of talks on Berlin by the four Allied occupying powers. They reached agreement in August 1971, and sanctioned talks began

Brandt's *Ostpolitik*

with the East Germans to improve facilities for communications between West Germany and West Berlin. In 1972 the two German states initialed a pact that made possible formal relations and cooperation; as a result, families were reunited, visits allowed, prisoners released, and relations formalized between the two Germanys. Relations deteriorated again in 1980 when the East German government increased the minimum currency-exchange requirement for Western visitors and again demanded full diplomatic recognition by West Germany.

In 1972 the Brandt government, by asking for a vote of confidence and then making ministers abstain from a vote, managed to bring about a dissolution of the Bundestag, even though this was, strictly speaking, an unconstitutional procedure. Although, in the general elections that followed, the Social Democrats scored their greatest election success, Brandt's policies of domestic reform later ran into severe difficulties caused in part by resistance from the Free Democrats and the impact of an economic recession. Unrest grew in the left wing of the Social Democratic Party and among university students. The ugly face of terrorist activities created a tense domestic political climate and caused a shift of public opinion toward more conservative views. Brandt's high prestige melted away, and in May 1974 he resigned after accepting responsibility for negligence in a spy scandal involving one of his personal aides.

Chancellorship of Helmut Schmidt

On May 16, 1974, Brandt was succeeded by Helmut Schmidt, who inherited a difficult political situation. The oil crisis had resulted in a worldwide economic recession, and the West Germans found to their dismay that the long period of steady economic growth they had enjoyed since the late 1950s appeared to have come to an end. Schmidt emerged as a courageous political leader who did not hesitate to make the stability of the economy his first objective. Together with President Valéry Giscard d'Estaing of France, he forged the European Monetary System, by which the enormous fluctuations of currencies could be halted and an element of stability brought into the international economy.

In foreign affairs, Schmidt soon arrived at the conclusion that improvements in East-West relations could only be attained by maintaining a determined security policy that reacted in kind against the Soviet rearmament policies of the day. He resolutely supported NATO policies of rearmament in Europe as a corollary to U.S.-Soviet agreements on the limitation of intercontinental nuclear missiles. Steadfast support of the Western alliance associated with constructive critique of Western, and in particular American, policies vis-à-vis the Soviet Union earned the Schmidt government high international regard. This regard was reflected in domestic politics; in the general elections of September 1980, the social-liberal coalition maintained control of the Bundestag.

The arrival in 1980 of a new ecology party, the Greens, to the left of the Social Democrats, significantly changed West Germany's political climate. The left wing of the Social Democrats became dissatisfied with the allegedly conservative style of existing governmental policies and demanded that government act upon a limitation of armaments, the abandonment of nuclear weaponry in the European theatre, the limitation of the use of nuclear energy for civil purposes, and the preservation of the environment. The firm stand taken by the social-liberal coalition on a policy of limited rearmament associated with the negotiations on nuclear disarmament brought Schmidt increasingly in conflict with his own party, which wanted to ban nuclear weapons from the European continent and at times even played with the idea of unilateral disarmament. The domestic political basis of the social-liberal coalition was also eroded because of the means by which the government tried to stem the tide of terrorist activities, which had reached a peak in 1975. But the final crisis of the Schmidt government developed over economic policies. The worldwide recession of the 1970s had convinced the right wing of the Free Democratic Party that a policy of economic retrenchment and return to the principles of the market economy was unavoidable. On Oct. 1, 1982—in a complicated procedure considered by some jurists to be unconstitutional—Schmidt was brought down by a vote of no confidence, and Helmut Kohl was installed as chancellor. Kohl's leadership of a conservative-liberal coalition continued after elections in 1983, in which Richard von Weizsäcker was chosen as president and head of state.

The conservative-liberal coalition (1982–). The Kohl government promised to bring about a general swing toward policies based upon firm allegiance to the Western alliance and to the principles of the market economy. Conservative politicians pleaded that the Germans ought to make peace with their own history and to reassert their own national identity, which had been overcast by an all-too-dominant concentration upon Germany's misdeeds under National Socialism. Kohl also attempted to put an end to the recrimination of the Germans for their past by participating in such events as the symbolic meeting with U.S. President Ronald Reagan at a military cemetery near Bitburg. Such events triggered a passionate debate among German historians, many of whom emphasized that any reinterpretation of German history must not tamper with the close affiliation with Western liberal tradition that had been achieved in the first three decades after World War II.

Chancellorship of Helmut Kohl

It was with considerable domestic political difficulties that the Kohl government got a moderate reform program under way, designed to strengthen the economy through a reduction of taxes and other indirect restrictions of economic activity. The elections of January 1987 confirmed that the electorate was not in favour of a radical reconstruction of the social system in neoconservative terms. The Kohl government did attempt to implement a major reform of the system of taxation in order to relieve the tax burden, particularly on the middle and upper-middle classes and on commerce and industry, while increasing some indirect taxation. The government also attempted a thoroughgoing reform of the social insurance system in the face of ever-increasing costs. A repatriation program was devised that made it advantageous for some of West Germany's foreign workers to return to their native countries. By 1985 about one-fifth of the Turkish workers reportedly had accepted grants and left the country.

Ostpolitik was continued in order to further ease the contacts between West and East Germans, primarily by concluding agreements on practical matters. A visit of the chairman of the Council of State of the German Democratic Republic, Erich Honecker, to the Federal Republic in September 1987 failed to bring any material results.

(W.J.M./Ed.)

THE GERMAN DEMOCRATIC REPUBLIC

In the Soviet Union's zone of occupation, a Soviet Military Administration assumed authority in June 1945 and, with the assistance of German communists, set about effecting a radical economic and social transformation. Former Nazis were summarily removed from the civil service and the educational and judicial systems. Private banks and many businesses were socialized without compensation, and large fortunes were confiscated. Landed estates in excess of 250 acres (100 hectares), comprising about a third of the zone's agriculture, were expropriated from some 3,000 landlords and divided among more than half a million farm workers, smallholders, and refugees. Heavy reparations were extracted by the Soviets, who dismantled whole factories, along with much rolling stock and even some rails, for reconstruction of areas of the Soviet Union that had been devastated by the German army.

Early political parties

In the summer of 1945, the Soviets permitted the revival of the Communist Party and the Social Democratic Party as well as the formation of two new parties, a Christian Democratic Union and a Liberal Democratic Party. Three years later a Democratic Farmer's Party and a National Democratic Party, the latter designed to appeal to former Nazis, were also approved. In April 1946, under pressure from the occupation authorities, the Social Democrats in the Soviet zone agreed to merge with the Communists, a step denounced by the Social Democrats in the West. The resulting Socialist Unity Party of Germany (Sozialistische Einheitspartei Deutschlands; SED) swept to victory with the ill-concealed aid of the Soviet occupation authorities in the first elections for local and regional representative

assemblies. But in October 1946, when elections were held under fairer conditions in Berlin, which was under four-power occupation, the Social Democrats tallied more than twice as many votes as the SED. Thereafter, the SED, which increasingly came under communist domination, avoided free, competitive elections.

In response to the steps that led in the West to the formation of the Federal Republic of Germany (West Germany), the SED began to establish parallel institutions in the Soviet zone. In the place of a democratically elected representative body, a self-proclaimed People's Congress for Unity and Just Peace, consisting of some 1,500 delegates chosen by the SED and organizations friendly to it—including some in the Western zones—convened in Berlin in late 1947. It established a smaller body of 330 in 1949 to draft a constitution for all of Germany. When it became clear that a West German government would be established, a so-called election for a new People's Congress was held in the Soviet zone in May 1949. Instead of choosing among candidates, voters were allowed only the choice of approving or rejecting—usually in less-than-secret circumstances—so-called unity lists of candidates drawn from all parties, as well as representatives of mass organizations controlled by the SED. By ensuring that communist candidates predominated in these unity lists, the SED determined in advance the composition of the new People's Congress. According to the official results, about two-thirds of the voters cast affirmative ballots. In subsequent elections, favourable margins in excess of 99 percent were routinely announced.

Formation of the new state. In October 1949, following the formation of the Federal Republic, the draft constitution went into effect in the Soviet zone, which became the German Democratic Republic (East Germany). The People's Congress was renamed the People's Chamber (Volkskammer) and, jointly with a second chamber composed of officials of the five federal states of the Soviet zone (which were abolished in 1952 in favour of a fully centralized administration), it designated the communist Wilhelm Pieck of the SED as president of East Germany on Oct. 11, 1949. The next day, the Volkskammer installed the former Social Democrat Otto Grotewohl as premier at the head of a Cabinet that was nominally responsible to the chamber. Decisive power lay, however, with the veteran communist functionary Walter Ulbricht. He held only the obscure position of deputy premier, but his post as general secretary (later first secretary) of the SED gave him the same sort of control over the party personnel and the government that had enabled Joseph Stalin to become dictator of the Soviet Union. In East Germany, as in the Soviet Union, the government served as the agent of an all-powerful communist-controlled party, which was in turn ruled from above by a Soviet-style, self-selecting Politburo.

The power of the party

The SED regime accorded priority to construction of a viable economy in what had previously been only part of Germany, and one that lacked rich natural resources. It used its centralized control over a planned economy to invest heavily in the construction of basic industries at the cost of the production of consumer goods. Reparations required that much productive capacity be diverted to Soviet needs. Despite an impressive rate of industrial growth, the standard of living thus remained low, lagging far behind that of West Germany. Even food was a problem, as thousands of farmers fled to the West each year rather than give in to mounting pressure to merge their land into the collective farms favoured by the communist leadership. Food rationing had to be continued long after it was ended in the West. The resulting material hardships, along with relentless ideological indoctrination, imprisonment of dissenters, and harassment of churches by a militantly atheistic regime, led many thousands of East Germans to flee to the West every year. In 1952 East Germany sealed its border with West Germany, but East Germans continued to leave through Berlin, where free movement still prevailed between the occupation sectors of the four-power city.

Unrest and its repercussions. Dissatisfaction with the government led to the first popular uprising in the post-war Soviet bloc when workers in East Berlin, the seat of the German Democratic Republic's government, went on strike on June 17, 1953, to protest against increased production quotas. When the regime failed to respond, the workers took to the streets and demanded a change in government. The rebellion quickly spread throughout East Germany and was quelled only when Soviet troops intervened, killing at least 21 and wounding hundreds of others. A wave of retribution followed, as some 1,300 were sentenced to prison for taking part in the uprising, which the East German government portrayed as a plot on the part of West Germany and the United States.

In the wake of the uprising, Stalin's successors in the Soviet Union upgraded the status of the German Democratic Republic. In 1954 reparations to the Soviet Union were halted, and Moscow proclaimed East Germany a sovereign state. In 1955 East Germany became a charter member of the Warsaw Pact, the Soviet bloc's military alliance. The SED leadership loosened ideological controls on artistic and intellectual activities somewhat, and the production of consumer goods was increased. Pressure on farmers to enter collective farms was relaxed. Agricultural yields improved, and the last food rationing was ended in 1958. Within a few years, however, the regime resumed its repressive measures and again shifted its economic priorities to favour the collectivization of agriculture and investment in industrial development at the cost of consumer goods. The flight of refugees through Berlin continued.

A brief relaxation

The Berlin Wall. In 1961 the flow of refugees to the West increased dramatically, bringing the total who had fled since the end of the war to some three million. On Aug. 13, 1961, the East German government surprised the world by sealing off East Berlin from West Berlin, first by barbed wire, later by construction of a concrete wall through the middle of the city. No one was thenceforth permitted to go to the West through the tightly guarded checkpoints without official permission, which was rarely granted. East Germans who sought to escape by climbing over the wall or tunneling under it risked death for the crime of "flight from the Republic." By stanching the hemorrhage of refugees, the regime stabilized the economy of the German Democratic Republic, which became the most prosperous of the Soviet bloc but nevertheless continued to lag far behind that of West Germany in both the quantity and quality of consumer goods. Ulbricht, who had assumed the powers of the presidency as head of a newly created Council of State upon President Pieck's death in 1960, imposed a new constitution in 1968 that sharply curtailed civil and political rights. He fell from power in 1971, however, after resisting the Soviet Union's policy of improving relations with the Federal Republic. His successor, Erich Honecker of the SED, quickly established tight control over the German Democratic Republic.

Relations with the West. In 1972 the Federal Republic entered into a treaty with the German Democratic Republic in line with the new *Ostpolitik* ("eastern policy") of West German Chancellor Willy Brandt's Cabinet. The two Germanys agreed to exchange permanent diplomatic missions, although West Germany declined to regard East Germany as a foreign country and so withheld full ambassadorial representation. Other Western governments, including that of the United States, soon established full diplomatic relations with the German Democratic Republic, which was admitted to the United Nations, along with the Federal Republic, in 1973. Spurred by large-scale West German credits, trade between the two Germanys increased. Eased travel restrictions permitted visits, first by West Germans to the East, then later, in the 1980s, by East Germans to the West. In 1981 Chancellor Helmut Schmidt paid an official visit to East Germany and, after numerous postponements, in 1987 SED chief Honecker was officially received in West Germany by Chancellor Helmut Kohl. Although Honecker had in 1974 effected revisions in the East German constitution that emphasized its irrevocable ties to the Soviet Union, tensions developed with Moscow when the SED reacted coolly to the reforms of Soviet leader Mikhail Gorbachev. In late 1988 the German Democratic Republic went so far as to refuse to permit distribution of a Soviet magazine that the SED denounced as subversive. (H.A.T.)

Germany reunited

Events in eastern Europe in 1989 set a new course for German-German relations. An easing of restrictions on emigration at midyear failed to accommodate the large number of East Germans attempting to leave the country. The opening to the West of other eastern European borders channeled an increasing exodus from the German Democratic Republic, which the government moved to stem through diplomatic as well as security measures. In the atmosphere of heightening tension that surrounded the republic's 40th anniversary ceremonies in early October, visiting Soviet leader Gorbachev stressed the urgent need for reform.

An organized political opposition, largely led by a group of intellectuals known as New Forum, staged massive peaceful demonstrations throughout the fall in Leipzig, Berlin, and other cities. Erich Honecker was ousted in October and succeeded by Egon Krenz, who resigned two months later. Following him were the brief interim tenures of Manfred Gerlach, East Germany's first noncommunist leader, and of Sabine Bergman-Pohl. Hans Modrow, the Dresden party chairman named premier in November, pursued a reformist policy.

By year's end, after a series of major government concessions, virtually all demands for liberalization had been met, including the relinquishing of the communist monopoly of power, the public investigation of party and government corruption, the announced disbanding of the security police, the lifting of travel restrictions, and, most dramatically, the opening of the Berlin Wall on November 9. Despite these developments, the tide of emigration continued, placing a burden on the West German economy that strained newly improved relations between the two countries.

As East Germans prepared for the first free elections in their nation's history, the issue of reunification with West Germany—specifically the timetable for such a process, which was now widely regarded both inside and outside German borders as inevitable—became the central point of contention among burgeoning political parties. The Alliance for Germany, an amalgam of parties advocating rapid reunification and supported by West German Chancellor Kohl's Christian Democratic Union, won a decisive victory at the polls on March 18, 1990. Also in March, representatives from the two Germanys as well as the United States, the Soviet Union, Britain, and France (the World War II Allied powers) convened in Bonn to begin charting the course toward a united Germany. The so-called grand coalition government, which assumed power in April under Premier Lothar de Maizière, held three-fourths of the seats in the Volkskammer (Parliament), a sufficient number to implement its proposed changes in the East German constitution.

Monetary union

A treaty signed in May established the terms of monetary union between the two Germanys—considered a prerequisite for political union. Despite initial resistance from West Germany's Social Democratic Party and opposing votes from the Greens and East Germany's Party of Democratic Socialism (the former communist party), the agreement was ratified by the respective German parliaments in June and became effective July 1. Under the administration of the Federal Republic's Bundesbank, the West German deutsche mark replaced the East German mark as that country's unit of currency, with a graduated exchange rate beginning at one-to-one. The Democratic Republic thereby abandoned its centrally planned economy in favour of the market-based West German system. The merger also brought the eastern sector under the Federal Republic's tax, social security, labour, and banking laws and abolished all border restrictions between the two countries.

While economic union rendered the course toward full union effectively irreversible, two formidable obstacles remained: the questions of military alliance and the German-Polish border. Responsibility regarding these issues rested in part with the Allied powers, which had engineered the German partition. The Federal Republic and its Western allies favoured NATO membership for a united Germany, while the Soviet Union, faced with the new uncertainties of the Warsaw Pact, first proposed nonalignment and demilitarization and subsequently advocated joint membership in the Eastern and Western organizations.

Talks between Chancellor Kohl and President Gorbachev produced an agreement on July 16 approving Germany's membership in NATO. In addition, the accord delineated terms for the withdrawal of Soviet troops from the eastern sector and the reduction of combined German troop strength, both over a three- to four-year period. A united Germany would sign the nuclear nonproliferation treaty and would ban the production or possession of all unconventional weapons. Just one day after the Soviet-German breakthrough, representatives from the two Germanys, the four Allied powers, and Poland reached an agreement guaranteeing Poland's existing borders. Freed of these final impediments, the members of the East German Volkskammer began debating the date for their country's dissolution.

A state treaty securing the political and social terms for reunification was signed by representatives of East and West Germany on August 31. The document reinstated Berlin as the German capital and scheduled formal union for Oct. 3, which was designated a national holiday. On September 12, officials from the two Germanys and the four Allied powers signed a treaty whereby the latter relinquished all rights of occupation and attendant responsibilities with regard to the former, thus ensuring united Germany's full sovereignty over its domestic and foreign affairs. In advance of ratification, the Allies agreed to implement these terms two days prior to reunification. On Oct. 3, 1990, less than one year after the opening of the Berlin Wall, German partition officially ended. Jubilant celebrations—an estimated one million revelers gathered on Berlin's Unter den Linden—gave way to the sober business of economic integration, social adjustment, and nationwide elections that faced the new Germany. (Ed.)

The end of partition

For later developments in the history of Germany, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 921, 923, 961, 963, and 972, and the *Index.*

BIBLIOGRAPHY

Physical and human geography: WALTER LAQUEUR, *Germany Today: A Personal Report* (1985), is a survey work on the former West Germany. Other general works include KARL RÖMER (ed.), *Facts About Germany: The Federal Republic of Germany,* 6th rev. ed. (1988); and RICHARD F. NYROP (ed.), *Federal Republic of Germany: A Country Study,* 2nd ed. (1982). Human geography is discussed in M.T. WILD, *West Germany: A Geography of Its People* (1979); LUIGI BARZINI, *The Impossible Europeans* (1983), discussing the nation in the context of the European federation; and GEORGE BAILEY, *Germans: The Biography of an Obsession* (1972). For statistics, see the yearbooks *Statistisches Jahrbuch für die Bundesrepublik Deutschland* and *Jahrbuch der Bundesrepublik Deutschland.* The West German economy is treated in HELMUT BÖHME, *An Introduction to the Social and Economic History of Germany* (1978; originally published in German, 6th ed., 1976); S.F. FROWEN, A.S. COURAKIS, and M.H. MILLER (comps.), *Monetary Policy and Economic Activity in West Germany* (1977); JEREMY LEAMAN, *The Political Economy of West Germany, 1945–85* (1988); GRANT KIRKPATRICK, *Employment, Growth, and Economic Policy* (1987); M.T. WILD (ed.), *Urban and Rural Change in West Germany* (1983); and *German Yearbook on Business History.*

The administrative and political structure of the former West Germany is treated in PETER J. KATZENSTEIN, *Policy and Politics in West Germany: The Growth of a Semi-Sovereign State* (1987); THOMAS ELLWEIN, *Das Regierungssystem der Bundesrepublik Deutschland,* 5th ed. (1983); MARION DÖNHOFF, *Foe into Friend: The Makers of the New Germany from Konrad Adenauer to Helmut Schmidt* (1982; originally published in German, 1981); and IAN DERBYSHIRE, *Politics in West Germany: From Schmidt to Kohl* (1987). Political parties are discussed in GORDON SMITH, *Democracy in Western Germany: Parties and Politics in the Federal Republic,* 3rd ed. (1986); ELIM PAPADAKIS, *The Green Movement in West Germany* (1984); and CHARLENE SPRETNAK and FRITJOF CAPRA, *Green Politics,* rev. ed. (1986). CHARLES BURDICK, HANS-ADOLF JACOBSEN, and WINFRIED KUDSZUS (eds.), *Contemporary Germany: Politics and Culture* (1984); and ARTHUR HEARNDEN, *Education, Culture,*

and Politics in West Germany (1976), are broad surveys. Intellectual life, arts, and culture are discussed in GORDON A. CRAIG, *The Germans* (1982); GARY D. STARK and BEDE KARL LACKNER (eds.), *Essays on Culture and Society in Modern Germany* (1982); *Cultural Life in the Federal Republic of Germany: A Survey,* trans. from German (1981), published by Inter Nationes; K. STUART PARKES, *Writers and Politics in West Germany* (1986); and JOHN SANDFORD, *The New German Cinema* (1980). (G.H.K.)

Comprehensive works on the former East Germany include STEPHEN R. BURANT (ed.), *East Germany: A Country Study,* 3rd ed. (1988); HARTMUT ZIMMERMAN (ed.), *DDR Handbuch,* 3rd ed. (1985), which was published in West Germany; and *Deutsche Demokratische Republik: Handbuch,* 2nd ed. (1984), published in East Germany. See also H. KOHL, J. MARCINEK, and B. NITZ, *Geography of the German Democratic Republic* (1986; originally published in German, 3rd ed., 1980); and A. ZIMM (ed.), *Berlin und sein Umland* (1988). Human geography is presented in HENRY KRISCH, *The German Democratic Republic: The Search for Identity* (1985); and TIMOTHY GARTON ASH, *"Und willst du nicht mein Bruder sein . . .": Die DDR heute* (1982), a book of broad observations focusing on particular cities and localities. EBERHARD SCHNEIDER, *The G.D.R.: The History, Politics, Economy, and Society of East Germany* (1978; originally published in German, 1975); and DAVID CHILDS, *The GDR: Moscow's German Ally,* 2nd ed. (1988), cover a broad spectrum of topics. See also *Information G.D.R.: The Comprehensive and Authoritative Reference Source of the German Democratic Republic,* 2 vol. (1989), compiled by the GDR Academy of Sciences. Economic policies and conditions are explored in IAN JEFFRIES and MANFRED MELZER (eds.), *The East German Economy* (1987); REINHARD POHL (ed.), *Handbook of the Economy of the German Democratic Republic* (1979; originally published in German, 1977); and *GDR and Eastern Europe: A Handbook* (1989; originally published in German, 1981). Statistical information is found in *Statistisches Jahrbuch der Deutschen Demokratischen Republik.*

The administrative, political, and social structure of the former East Germany is studied in MIKE DENNIS, *German Democratic Republic: Politics, Economics, and Society* (1988); C. BRADLEY SCHARF, *Politics and Change in East Germany: An Evaluation of Socialist Democracy* (1984); G.E. EDWARDS, *GDR Society and Social Institutions: Facts and Figures* (1985); and MARGRETE SIEBERT KLEIN, *The Challenge of Communist Education: A Look at the German Democratic Republic* (1980). A survey of intellectual life and popular culture is found in *Cultural Life in the GDR: Review and Current Trends* (1982), produced by Panorama DDR in the series "First-Hand Information." (T.H.El.)

Comparative studies: A large number of works tend to compare the two former states or to treat them together. T.H. ELKINS, *Germany,* 3rd ed. (1972); and ROY E.H. MELLOR, *The Two Germanies: A Modern Geography* (1978), are two such treatments. T.H. ELKINS, *Berlin: The Spatial Structure of a Divided City* (1988), deals with both sides of the wall. *Statistisches Jahrbuch Berlin* (annual) includes a summary of statistics on East Berlin that are otherwise not readily available. TERENCE PRITTIE, *My Germans, 1933–1983* (1983), explores social life and customs and national characteristics, as do JOHN ARDAGH, *Germany and the Germans: An Anatomy of Society Today* (1987); and HARRY G. SHAFFER, *Women in the Two Germanies: A Comparative Study of a Socialist and a Non-Socialist Society* (1981). For the economy, see HERBERT WILKENS, *The Two German Economies: A Comparison Between the National Product of the German Democratic Republic and the Federal Republic of Germany,* rev. ed. (1981; originally published in German, 1976). RALF DAHRENDORF, *Society and Democracy in Germany* (1967, reissued 1979; originally published in German, 1965), surveys social and political characteristics. Relations between the two countries are the subject of LAWRENCE L. WHETTEN, *Germany East and West: Conflicts, Collaboration, and Confrontation* (1980); and ERNEST D. PLOCK, *The Basic Treaty and the Evolution of East–West German Relations* (1986). German literature and culture are looked at in LOWELL A. BANGERTER, *German Writing Since 1945: A Critical Survey* (1988); and MARION ADAMS (ed.), *The German Tradition: Aspects of Art and Thought in the German-Speaking Countries* (1971). Comparative histories include HENRY ASHBY TURNER, JR., *The Two Germanies Since 1945* (1987), the only dual history; ALFRED GROSSER, *Germany in Our Time: A Political History of the Postwar Years* (1971; originally published in French, 1970); and V.R. BERGHAHN, *Modern Germany: Society, Economy, and Politics in the Twentieth Century,* 2nd ed. (1987). The question of reunification is explored in RENATA FRISCH-BOURNAZEL, *Confronting the German Question: Germans on the East–West Divide* (1988; originally published in German, 1986); and RONALD A. FRANCISCO and RICHARD L. MERRITT, *Berlin Between Two Worlds* (1986). (G.H.K./T.H.El.)

History. *Ancient Germany:* A good, well-illustrated general account of the early period is given in HERBERT SCHUTZ, *The Prehistory of Germanic Europe* (1983). TIM CORNELL and JOHN MATTHEWS, *Atlas of the Roman World* (1982), clarifies with maps and illustrations the social and economic background of Roman-German interaction. Other useful analyses of the early period include E.A. THOMPSON, *The Early Germans* (1965); ROLF HACHMANN, *The Germanic Peoples* (1971; originally published in German, 1971); and MALCOLM TODD, *The Northern Barbarians, 100 BC–AD 300,* rev. ed. (1987). A comprehensive study of the later period is LUCIEN MUSSET, *The Germanic Invasions: The Making of Europe, AD 400–600* (1975; originally published in French, 1965). Individual peoples are discussed in EDWARD JAMES, *The Franks* (1988); and JAMES CAMPBELL, ERIC JOHN, and PATRICK WORMALD, *The Anglo-Saxons* (1982). (P.J.He.)

Merovingians and Carolingians: PATRICK J. GEARY, *Before France and Germany: The Creation and Transformation of the Merovingian World* (1988), examines the period from the perspective of the unity of late antiquity. A general survey of Carolingian Europe, with an emphasis on cultural history, is presented in ROSAMOND MCKITTERICK, *The Frankish Kingdoms Under the Carolingians, 751–987* (1983). Frankish society, its politics, religion, and ecclesiastical organization, are discussed in J.M. WALLACE-HADRILL, *The Long-Haired Kings, and Other Studies in Frankish History* (1962, reprinted 1982), and *The Frankish Church* (1983). Important essays on early German and Frankish society are collected in TIMOTHY REUTER (ed.), *The Medieval Nobility: Studies on the Ruling Classes of France and Germany from the Sixth to the Twelfth Century* (1979). (P.J.G.)

Medieval Germany to 1250: G. BARRACLOUGH, *The Origins of Modern Germany,* 3rd ed. (1988), is a good analysis of medieval German history. General surveys include KARL HAMPE, *Germany Under the Salian and Hohenstaufen Emperors* (1973; originally published in German, 12th ed., 1968), an important and well-organized, though old-fashioned, narrative history; HORST FUHRMANN, *Germany in the High Middle Ages, c. 1050–1200* (1986; originally published in German, 2nd ed., 1983), covering such fundamentals as economy and cultural horizons; and ALFRED HAVERKAMP, *Medieval Germany, 1056–1273* (1988; originally published in German, 1984), providing an informative though mainly German bibliography. On the 10th and early 11th centuries, see JOSEF FLECKENSTEIN, *Early Medieval Germany* (1978; originally published in German, 1974). K.J. LEYSER, *Rule and Conflict in an Early Medieval Society: Ottonian Saxony* (1979, reprinted 1989), explores the main forces of development of the Saxon empire and its society from 900 to 1024. The same author's *Medieval Germany and Its Neighbours, 900–1250* (1982), studies warfare, the nobility, and German-Byzantine and German-English relations. See also his "Crisis of Medieval Germany," in *Proceedings of the British Academy,* vol. 69 (1983), pp. 410–443, on the Saxon rising of 1073. BENJAMIN ARNOLD, *German Knighthood, 1050–1300* (1985), is a fine study of knights and *ministeriales.* Comprehensive surveys of the 11th–13th centuries include FRIEDRICH PRINZ, *Grundlagen und Anfänge, Deutschland bis 1056* (1985); and HAGEN KELLER, *Zwischen regionaler Begrenzung und universalem Horizont: Deutschland im Imperium der Salier und Staufer, 1024 bis 1250* (1986). (K.J.L.)

From 1250 to 1493: JOACHIM LEUSCHNER, *Germany in the Late Middle Ages* (1980; originally published in German, 1975), provides a basic introduction to political history, which should be supplemented by the more complex F.R.H. DU BOULAY, *Germany in the Later Middle Ages* (1983); and by LAWRENCE G. DUGGAN, "Representative Assemblies, German," in *Dictionary of the Middle Ages,* vol. 10 (1988), pp. 328–334. On regional and local developments, see F.L. CARSTEN, *The Origins of Prussia* (1954, reprinted 1981), and *Princes and Parliaments in Germany, from the Fifteenth to the Eighteenth Century* (1959, reprinted 1963); OTTO BRUNNER, *Land und Herrschaft: Grundfragen der territorialen Verfassungsgeschichte Österreichs im Mittelalter,* 5th ed. (1965, reprinted 1984); HENRY J. COHN, *The Government of the Rhine Palatinate in the Fifteenth Century* (1965); PHILIPPE DOLLINGER, *The German Hansa* (1970; originally published in French, 1964); LAWRENCE G. DUGGAN, *Bishop and Chapter: The Governance of the Bishopric of Speyer to 1552* (1978); and THOMAS A. BRADY, JR., *Turning Swiss: Cities and Empire, 1450–1550* (1985). Useful documentary sources are collected in the first, and studies in the second, of the following: GERALD STRAUSS (ed.), *Manifestations of Discontent in Germany on the Eve of the Reformation* (1971), and *Pre-reformation Germany* (1972). HERMANN KELLENBENZ, *Deutsche Wirtschaftsgeshichte,* vol. 1: *Von den Anfängen bis zum Ende des 18. Jahrhunderts* (1977), surveys economic history. For cultural history, see WILLY ANDREAS, *Deutschland vor der Reformation: Eine Zeitenwende,* 7th ed. (1972); and the more up-to-date HANS-FRIEDRICH ROSENFELD

and HELLMUT ROSENFELD, *Deutsche Kultur im Spätmittelalter 1250–1500* (1978). A fine overview of intellectual life is found in the collection of GERHART HOFFMEISTER (ed.), *The Renaissance and Reformation in Germany* (1977); and on the extraordinary flowering of German sculpture there are two masterly works by MICHAEL BAXANDALL, *South German Sculpture, 1480–1530* (1974), and *The Limewood Sculptors of Renaissance Germany* (1980). (L.G.D.)

From 1493 to c. 1760: HAJO HOLBORN, *A History of Modern Germany: The Reformation* (1959, reprinted 1982), provides a comprehensive general account, with emphasis on political events. A.G. DICKENS, *The German Nation and Martin Luther* (1974), explores the main features of German society around 1500 through the analysis of the widespread influence of Luther. HEIKO A. OBERMAN, *Luther: Man Between God and the Devil* (1989, originally published in German, 1982), is a psychologically compelling biography of the reformer. PETER BLICKLE, *The Revolution of 1525: The German Peasants' War from a New Perspective* (1981; originally published in German, 1975), stresses the common grievances that brought peasants and the urban poor together in what the author sees as "the revolution of the common man." KARL BRANDI, *The Emperor Charles V: The Growth and Destiny of a Man and of a World-Empire* (1939, reissued 1980; originally published in German, 1937), presents Charles V as the protagonist of an ambitious plan to make the medieval idea of empire work in the early modern period. JAMES A. VANN and STEVEN W. ROWAN (eds.), *The Old Reich: Essays on German Political Institutions, 1495–1806* (1974), illustrates the successes and failures of the empire as a constitutional system. MERRY E. WIESNER, *Working Women in Renaissance Germany* (1986), offers a pioneering examination of the changing situation of women engaged in trades and professions between 1500 and 1600. GEOFFREY PARKER (ed.), *The Thirty Years' War* (1984), surveys the military, political, social, and cultural aspects of the war. JOHN G. GAGLIARDO, *Reich and Nation: The Holy Roman Empire as Idea and Reality, 1763–1806* (1980), looks at the empire's efforts and failure to survive in the age of the French Revolution. R.J.W. EVANS, *The Making of the Habsburg Monarchy, 1550–1700: An Interpretation* (1979, reprinted 1984), analyzes the events leading to the construction of the Austrian empire in central and eastern Europe. MACK WALKER, *German Home Towns: Community, State, and General Estate, 1648–1871* (1971), presents a vivid picture of the development of the small and middle-sized German town as a distinct political and cultural environment. RUDOLF VIERHAUS, *Germany in the Age of Absolutism* (1988; originally published in German, 1978), discusses the German principalities in the aftermath of the wars against Louis XIV. (Ge.St.)

From c. 1760 to 1871: For a good account of central Europe during the period of the Enlightenment, see HENRI BRUNSCHWIG, *Enlightenment and Romanticism in Eighteenth-Century Prussia* (1974; originally published in French, 1947), an interpretive socioeconomic history; and C.B.A. BEHRENS, *Society, Government, and the Enlightenment: The Experiences of Eighteenth-Century France and Prussia* (1985), a comparative history of the events and developments that led to revolution in France and stabilization in Germany. HANS ROSENBERG, *Bureaucracy, Aristocracy, and Autocracy* (1958), describes the nature of authority. The transformation of civic ideals is the theme of FRIEDRICH MEINECKE, *Cosmopolitanism and the National State* (1970; originally published in German, 1908); his *The Age of German Liberation, 1795–1815* (1977; originally published in German, 1906), is an excellent concise political history of the Prussian state and the main figures of the period. The prevailing attitudes in the Holy Roman Empire toward the fall of the ancien régime are examined by G.P. GOOCH, *Germany and the French Revolution* (1920, reprinted 1966). T.C.W. BLANNING, *The French Revolution in Germany: Occupation and Resistance in the Rhineland, 1792–1802* (1983), is a reexamination of this period of German history from new positions. REINHOLD ARIS, *History of Political Thought in Germany from 1789 to 1815* (1936, reprinted 1965); and EUGENE N. ANDERSON, *Nationalism and the Cultural Crisis in Prussia, 1806–1815* (1939, reprinted 1966), depict the intellectual effects of the French hegemony. For political developments, see H.A.L. FISHER, *Studies in Napoleonic Statesmanship: Germany* (1903, reprinted 1969); and GUY S. FORD, *Stein and the Era of Reform in Prussia, 1807–1815* (1922, reissued 1965). ENNO KRAEHE, *Metternich's German Policy,* 2 vol. (1963–83), describes the War of Liberation; while HAROLD G. NICOLSON, *The Congress of Vienna* (1946, reprinted 1974), deals with the reconstruction of Europe. For the period of the German Confederation, see the vigorous but opinionated account by HEINRICH G. VON TREITSCHKE, *History of Germany in the Nineteenth Century,* 7 vol. (1915–19, reissued 1968; originally published in German, 5 vol., 1879–89). THEODORE S. HAMEROW, *Restoration, Revolution, Reaction* (1958, reprinted 1966), stresses the connection between politics and economics. JOHN H. CLAPHAM, *The Economic Development of France and Germany, 1815–1914,* 4th ed. (1936, reprinted 1968); and WILLIAM O. HENDERSON, *The Zollverein,* 3rd ed. (1984), examine the transformation of the economy of central Europe. VEIT VALENTIN, *1848: Chapters of German History* (1940, reprinted 1965; originally published in German, 1930), emphasizes the political events of the revolution; WERNER E. MOSSE, *The European Powers and the German Question, 1848–71* (1958, reprinted 1981), looks at the diplomatic aspects. The achievement of national unification is described from the Prussian point of view by HEINRICH VON SYBEL, *The Founding of the German Empire by William I,* 7 vol. (1890–98, reprinted 1968; originally published in German, 1889–95); while HEINRICH FRIEDJUNG, *The Struggle for Supremacy in Germany, 1859–1866* (1935, reprinted 1966; originally published in German, 10th ed., 2 vol., 1916–17), looks at it from the Austrian side. EUGENE N. ANDERSON, *The Social and Political Conflict in Prussia, 1858–1864* (1954, reissued 1976), deals with the constitutional struggle in the Hohenzollern kingdom. OTTO PFLANZE, *Bismarck and the Development of Germany: The Period of Unification, 1815–1871* (1963, reprinted 1971); and ERICH EYCK, *Bismarck and the German Empire,* 3rd ed. (1968), portray the architect of the new Germany and his impact on the European balance of power. The origins of the Franco-German War are traced in HERMANN ONCKEN, *Napoleon III and the Rhine: The Origin of the War of 1870–1871* (1928, reissued 1967; originally published in German, 1926). The intellectual life of 18th- and 19th-century Germany is discussed in HERMAN GLASER (ed.), *The German Mind of the Nineteenth Century: A Literary and Historical Anthology* (1981). (T.S.H.)

From 1871 to 1918: General historical surveys include HANS-ULRICH WEHLER, *The German Empire, 1871–1918* (1985; originally published in German, 5th ed., 1983), stressing the continuity between the empire and the Nazi era; and DAVID BLACKBOURN and GEOFF ELEY, *The Peculiarities of German History: Bourgeois Society and Politics in Nineteenth-Century Germany* (1984), emphasizing the similarities between German modernization and that of other European nations and the importance of industrial capitalism rather than the Junker elite in supporting the authoritarian character of the German Empire. Political analyses are presented in MARGARET LAVINIA ANDERSON, *Windthorst: A Political Biography* (1981), a major study of the Roman Catholic party leader and of Catholic politics; RICHARD J. EVANS (ed.), *Society and Politics in Wilhelmine Germany* (1978), a collection of essays on local politics; LOTHAR GALL, *Bismarck, the White Revolutionary,* 2 vol. (1986; originally published in German, 1980), a study of Bismarck's influence on imperial politics; and JAMES J. SHEEHAN, *German Liberalism in the Nineteenth Century* (1978, reprinted 1983), an excellent analysis of the activities of the two liberal parties on the local and national levels, with emphasis on their growing acceptance of the empire.

Foreign relations are studied in LAMAR CECIL, *The German Diplomatic Service, 1871–1914* (1976), a careful look at the background and education of the German diplomatic corps before World War I; V.R. BERGHAHN, *Germany and the Approach of War in 1914* (1973), a thorough account of German foreign policy, especially the influence of the navy and army; PAUL M. KENNEDY, *The Rise of the Anglo-German Antagonism, 1860–1914* (1980), an exhaustive work on the growing antagonism that led the countries to opposite sides in World War I; and FRITZ FISCHER, *Germany's Aims in the First World War* (1967; originally published in German, 3rd ed., 1964), a pioneering study of the culpability of German elites in the outbreak of World War I, stressing the continuity between the two world wars.

Economic developments of the period are surveyed in FRITZ STERN, *Gold and Iron: Bismarck, Bleichröder, and the Building of the German Empire* (1977, reprinted 1987), an excellent study, based on archival sources, of the role of Bismarck and his banker in the economy; KENNETH D. BARKIN, *The Controversy Over German Industrialization, 1890–1902* (1970), a discussion of the politico-economic debate on whether Germany should industrialize or remain a primarily agricultural country; and GERALD D. FELDMAN, *Army, Industry, and Labor in Germany, 1914–1918* (1966), a thorough study of the changes in the German economy during the four years of World War I.

Cultural and social issues are addressed in FRITZ STERN, *The Politics of Cultural Despair: A Study in the Rise of the Germanic Ideology* (1961, reprinted with a new preface, 1974), a study of intellectuals' disenchantment with modern industrial society and their hopes for a spiritual revolution. The history of German socialism is detailed in GARY P. STEENSON, *"Not One Man! Not One Penny!": German Social Democracy, 1863–1914* (1981), a comprehensive treatment of the Social Democratic Party from its origins to the war; VERNON L. LIDTKE, *The Outlawed Party: Social Democracy in Germany, 1878–1890* (1966),

an important study of the 12 years the party was banned; and JÜRGEN KOCKA, *Facing Total War: German Society, 1914–1918* (1984; originally published in German, 1973), an examination of the effects of World War I on different social groups. (Ke.B.)

From 1918 to 1945: A sophisticated, though somewhat dated, survey of modern German history is found in HAJO HOLBORN, *A History of Modern Germany: 1840–1945* (1969, reprinted 1982). DIETRICH ORLOW, *A History of Modern Germany: 1871 to Present* (1987), offers a good look at the difficult century through which Germany has just passed. A.J. RYDER, *The German Revolution of 1918: A Study of German Socialism in War and Revolt* (1967), treats both the democratic and revolutionary socialists in 1918 and 1919 along with their conflicting visions for change. Vol. 3 of H.W.V. TEMPERLEY (ed.), *A History of the Peace Conference of Paris,* 6 vol. (1920–24, reissued 1969), provides useful commentary and documentation on the peace settlement as related to Germany, with the text of the Treaty of Versailles in its entirety. An analysis of the peace settlement that emphasizes the fear of Bolshevik expansion as a factor shaping the various treaties is found in ARNO J. MAYER, *Politics and Diplomacy of Peacemaking: Containment and Counterrevolution at Versailles, 1918–1919* (1967). A still unsurpassed account of political developments in Germany during the period is S. WILLIAM HALPERIN, *Germany Tried Democracy: A Political History of the Reich from 1918 to 1933* (1946, reissued 1965). A stimulating introduction to the brilliance of intellectual and cultural creativity in Germany during the 1920s is provided in PETER GAY, *Weimar Culture: The Outsider as Insider* (1968, reprinted 1981). OTTO FRIEDRICH, *Before the Deluge: A Portrait of Berlin in the 1920s* (1972, reprinted 1986), is a highly readable discussion of the cultural and intellectual capital of Germany, heavily based upon interviews with important figures of the period. The best introduction to Nazism still is ALAN BULLOCK, *Hitler, a Study in Tyranny,* rev. ed. (1962). A comprehensive study of National Socialism emphasizing the totalitarian nature of the regime is available in KARL DIETRICH BRACHER, *The German Dictatorship: The Origins, Structure, and Effects of National Socialism* (1970; originally published in German, 1969). A useful companion to this work is MARTIN BROSZAT, *The Hitler State: The Foundation and Development of the Internal Structure of the Third Reich* (1981; originally published in German, 1969), which focuses upon the chaotic inefficiency the author believes characterized the Nazi political structure. The standard work on Nazi foreign policy is GERHARD L. WEINBERG, *The Foreign Policy of Hitler's Germany: Diplomatic Revolution in Europe, 1933–36* (1970, reprinted 1983), and *The Foreign Policy of Hitler's Germany: Starting World War II, 1937–1939* (1980). One of the best single-volume treatments of World War II is PETER CALVOCORESSI and GUY WINT, *Total War: The Causes and Courses of the Second World War,* 2nd rev. ed. (1989). RAUL HILBERG, *The Destruction of the European Jews,* rev. ed., 3 vol. (1985), remains the standard and most comprehensive study of the efforts by the Nazis to exterminate the Jewish people. German civilian life during the war is explored in EARL R. BECK, *Under the Bombs: The German Home Front, 1942–1945* (1986). (K.A.Sc.)

After 1945: Comprehensive historical surveys of West Germany include KARL DIETRICH BRACHER *et al.* (eds.), *Geschichte der Bundesrepublik Deutschland,* 5 vol. in 6 (1981–87); WOLFGANG BENZ (ed.), *Die Bundesrepublik Deutschland,* 3 vol. (1983); DIETRICH THRÄNHARDT, *Geschichte der Bundesrepublik Deutschland* (1986); MICHAEL BALFOUR, *West Germany: A Contemporary History* (1982); and RICHARD L. MERRITT and ANNA MERRITT (eds.), *Living with the Wall, 1961–1985* (1985). Analysis of the Allied occupation is found in ANDREAS HILLGRUBER, *Deutsche Geschichte 1945–1986: die "deutsche Frage" in der Weltpolitik,* 6th ed. (1987); and CHRISTOPH KLESSMANN, *Die doppelte Staatsgründung: deutsche Geschichte 1945–1955* (1982). For West German foreign relations, see CHRISTIAN HACKE, *Weltmacht wider Willen: die Aussenpolitik der Bundesrepublik Deutschland* (1988). GERHARD A. RITTER and MERITH NIEHUSS, *Wahlen in der Bundesrepublik Deutschland: Bundestags- und Landtagswahlen 1946–1987* (1987), gives statistics on elections. (W.J.M.)

MARTIN MCCAULEY, *The German Democratic Republic Since 1945* (1983), is a comprehensive historical survey. Special histories of narrower periods include GREGORY W. SANDFORD, *From Hitler to Ulbricht: The Communist Reconstruction of East Germany, 1945–46* (1983); JOHN DORNBERG, *The Other Germany* (1968), concentrating on the 1960s; and DAVID CHILDS (ed.), *Honecker's Germany* (1985). (H.A.T.)

Goethe

A German writer who is universally acknowledged to be one of the giants of world literature, Johann Wolfgang von Goethe was perhaps the last European to attempt the mastery and many-sidedness of the great Renaissance personalities: critic, journalist, painter, theatre manager, statesman, educationalist, natural philosopher. The bulk and diversity of his output is in itself phenomenal: his writings on science alone fill about 14 volumes. In the lyric vein he displayed a command of a unique variety of theme and style; in fiction he ranged from fairy tales, which have proved a quarry for psychoanalysts, through the poetic concentration of his shorter novels and *Novellen* (novellas) to the "open," symbolic form of *Wilhelm Meister;* in the theatre, from historical, political, or psychological plays in prose through blank-verse drama to his *Faust,* one of the masterpieces of modern literature. He achieved in his 82 years a wisdom often termed Olympian, even inhuman; yet almost to the end he retained a willingness to let himself be shaken to his foundations by love or sorrow. He disciplined himself to a routine that might armour him against chaos; yet he never lost the power of producing magical short lyrics in which the mystery of living, loving, and thinking was distilled into sheer transparency.

By courtesy of the Bayerische Staatsgemaldesammlungen, Munich, Neue Pinakothek



Goethe, oil painting by Joseph Karl Stieler, 1828. In the Neue Pinakothek, Munich.

And at the last there was granted him a gift, uncanny even to himself, of tapping at will the springs of creativity in order to complete the work he had carried with him for 60 years. When, a few months before his death, he sealed his *Faust,* he bequeathed it with ironic resignation to the critics of posterity to discover its imperfections. Its final couplet, "Das Ewig-Weibliche/Zieht uns hinan" ("Eternal Womanhead/Leads us on high"), epitomizes his own feeling about the central polarity of human existence: woman was to him at once man's energizer and his civilizer, source of creative life and focus of the highest endeavours of both mind and spirit.

There was in Goethe a natural, if not always painless, swing between poles of existence often thought to be mutually exclusive and an innate commitment to change and process. And, in the last letter he was to write, he rounded off what has sometimes been called his greatest work, his life, by setting the seal of his approval on a mode of growth that sees the art of living as the intensification of inborn talents through a judicious surrender to the natural rhythm of opposing tendencies.

Early life and influences. Born on August 28, 1749, in Frankfurt am Main, Goethe came of middle-class stock, the *Bürgertum* that he never ceased to praise as a breeding ground of the finest culture. His father, Johann Kaspar Goethe, was of north German extraction. A retired lawyer, he was able to lead a life of cultured leisure, travelling in Italy and amassing a well-stocked library and picture gallery in his handsomely furnished house. Goethe's mother, Katharine Elisabeth Textor, was the daughter of a *Bürgermeister* (mayor) of Frankfurt; she opened up to her son valued connections with the patriciate of the free city. Thus even in his heredity Goethe unites those opposing tendencies that have always prevailed in German lands: the intellectual and moral rigour of the north and the easygoing artistic sensuousness of the south. Of eight children, only Wolfgang, the firstborn, and his sister, Cornelia, survived.

In his autobiography, *Dichtung und Wahrheit* ("Poetry and Truth"), Goethe left an unforgettable picture of a happy childhood. Here are set out with acute psychological insight the emotional complexities of his bond with Cornelia, which found expression in numerous portrayals of the brother–sister relationship in his works; his passionate attachment to a barmaid, Gretchen, which foreshadowed the rejection pattern of many of his loves; the broadening of outlook that came with French occupation during the Seven Years' War; the coronation of Joseph II in the Frankfurt *Römer,* with its indelible impressions of medieval pageantry; and the fervent religiosity of Pietistic circles, which led him to declaim F.G. Klopstock's *Messias* ("Messiah") as a kind of Lenten exercise, to write a prose epic on Joseph and a poem on Christ's descent into hell. The French army had brought its own troupe of actors, and their performances intensified a passion for the stage, first kindled in him by his grandmother's gift of a puppet theatre, and inspired a lifelong devotion to Racine. A love of things English was fostered by friendship with a young clothier from Leeds (Goethe's paternal grandfather was a fashionable tailor) with whom Cornelia, seeing herself as the heroine of a Richardsonian novel, fell hopelessly in love. Wolfgang's reaction was the inception of a novel in letters, a kind of linguistic exercise in which four brothers correspond in different languages.

Study at Leipzig

In October 1765 Goethe was sent to study law at his father's old University of Leipzig, though he himself would have preferred to read classics in the newly founded university at Göttingen, where English influence prevailed. In Leipzig, or "little Paris" as he calls it in *Faust,* by contrast, a world of elegance and fashion made the young provincial feel like a fish out of water. The Frenchifying influence of the critic J.C. Gottsched still dominated the theatre and provided a repertory of the best plays of contemporary Europe. But C.F. Gellert, poet and author of fables and hymns, now in the heyday of his fame, presented the new sensibility of Edward Young, Laurence Sterne, and Samuel Richardson. Goethe praised Gellert's lectures as "the foundation of German moral culture" and learned from them invaluable lessons in epistolary style and in social conduct. Gellert's literary influence was reinforced by the robust elegance and ironic sagacity of the novels, tales, and epics of C.M. Wieland. Wieland's work was brought to Goethe's notice by A.F. Oeser, a friend and teacher of the archaeologist and art historian J.J. Winckelmann, who profoundly influenced European fashions in art. From Oeser, Goethe learned a love of Greek art and two things that stood him in good stead all his life: to use his eyes and to master the craft of whatever he undertook. A visit to Dresden, "the Florence of the north," as the poet and critic J.G. Herder called it, opened his eyes to the splendours of Rococo architecture as well as classical statuary. Nor was music neglected in his education; a new 18th-century concert society, under the direction of the musician and composer J.A. Hiller,

provided splendid performances, which became world famous as the *Gewandhaus* concerts.

Early works in the Rococo fashion

The literary harvest of Goethe's Leipzig period manifested itself in a songbook written in the prevailing Rococo mode—songs praising love and wine in the manner of the Greek poet Anacreon. Appropriately titled *Das Leipziger Liederbuch* (*The Leipzig Song Book*), it was ostensibly inspired by the daughter of the wine merchant at whose tavern he took his midday meal. But neither his 1766–67 poems *Das Buch Annette* ("The Book Annette"; as he called her in Rococo fashion) nor the *Neue Lieder* ("New Songs") of 1769 made any pretense of real passion. Yet it was in connection with these literary trifles that he subsequently made the famous and much abused statement that all his works were "fragments of a great confession." The same note is struck in two plays written in alexandrine verse (a 12-syllable iambic line borrowed from the French), *Die Laune des Verliebten* ("The Mood of the Beloved") and a more sombre farce, *Die Mitschuldigen* ("The Accomplices"), which foreshadows the psychological preoccupations of later works. From then on, Rococo was one element in Goethe's repertoire, to be drawn on as occasion demanded. It was to reappear in the setting of *Torquato Tasso* and *Die Wahlverwandtschaften* (*Elected Affinities*); he was to pay tribute to its charm in *Anakreons Grab* ("Anacreon's Grave"; 1806) and amalgamate it with Eastern influence in enchanting poems of the *West-östlicher Divan* ("Divan of East and West").

Works of the storm and stress period. Goethe's stay in Leipzig was cut short by severe illness, and by the autumn of 1768 he was back home. A long convalescence fostered introspection and religious mysticism. He played with alchemy, astrology, and occult philosophy, all of which left their mark on *Faust.* On his recovery it was decided that he should pursue legal studies in Strassburg as a first stage on the way to Paris and the Grand Tour (never actually completed). His stay there proved a turning point for his whole life and work. In this German capital of a French province, he experienced a reaction against the cosmopolitan atmosphere of Leipzig and under the impact of the great cathedral proclaimed his conversion to the Gothic German ideal. More decisive still was the influence of J.G. Herder, who spent the winter of 1770–71 there undergoing treatment for his eyes. From him Goethe learned the role played by touch, the haptic sense, in the growth of the mind; a new view of the artist as a creator fashioning forms expressive of feeling; a new theory of poetry as the original and most vital language of man; the virtues of a new style, that of the *Volkslied* (folk song) and the poetry of "primitive" peoples as enshrined in the Bible, the epics of Homer, and the poems attributed (falsely) to Ossian, a 3rd-century Celtic poet. It is this new sense of felt immediacy, and of the plasticity of his linguistic medium, that informs the lyrics Goethe wrote to one of his early loves, Friederike Brion, the pastor's daughter of Sesenheim. They mark the beginning of a new epoch in the German lyric. Such poems as "Mailied" ("May Song") and "Willkommen und Abschied" ("Welcome and Farewell") are still the most popular, though not the greatest, of his *Lieder.* The latter, especially in its revised form of 1790, touchingly expresses the guilt he felt that this time he himself had the role of deserter and rejecter, and the whole idyll as recounted in *Dichtung und Wahrheit* reveals that cross-fertilization of life and literature that he increasingly saw as a potent factor in human development.

Influence of J.G. Herder

If, as Herder maintained, energy was one of the marks of poetry, it was clearly in the passions acted out on the stage that it could find its most vital expression. And where more vital than in the colossal figures of the "Gothic Shakespeare"? In writing the *Geschichte Gottfriedens von Berlichingen mit der eisernen Hand dramatisiert* (1771; "The Dramatized History of Gottfried von Berlichingen of the Iron Hand"), Goethe was deliberately vying with Shakespeare. For the real Götz, who died two years before Shakespeare was born, was near enough in time to represent that bustling spacious 16th century, the animal vitality of which contrasted so forcibly with the straitlaced affectations of Goethe's own day. With the publication in 1773 of *Götz von Berlichingen,* a radically tautened version of that "History," the Shakespeare cult was launched, and the Sturm und Drang (storm and stress) movement was provided with its first major work of genius. The manifesto of the movement, heralded by Goethe's enthusiastic *Rede zum Schakespears Tag* ("Conversation from Shakespeare's Day"), had appeared after Goethe's return to Frankfurt in August 1771. "Von deutscher Art und Kunst" ("Concerning German Nature and Art"), as it was called, contained a defense of German nationality by the historian J.M. Möser, two essays by Herder championing Ossian and Shakespeare, and a rhapsody on Gothic architecture by Goethe.

The success of *Götz von Berlichingen*

Though ostensibly in practice as a lawyer, the young poet now found himself caught up in a whirl of literary and social duties—helping to edit the *Frankfurter Gelehrte Anzeigen* ("Frankfurt Scholarly Reviews"), for instance—and it was to break loose from this that he left for Wetzlar, seat of the supreme court of the Empire. But again literature won the day over law, and an impassioned yet self-ironic ode in free verse, "Wandrers Sturmlied" ("Wanderer's Storm Song"), is testimony both to a recently inspired admiration for Pindar, the greatest lyric poet of ancient Greece, and to a hesitant certainty that he himself might be destined for greatness. And in Wetzlar he experienced a new passion, this time for a girl safely out of reach from the start, Charlotte Buff. Her betrothed, Johann Christian Kestner, showed great understanding until, as it seemed to him, he found the affair exposed to public gaze in *Die Leiden des jungen Werthers* (*The Sorrows of Young Werther;* 1774).

But much besides the Wetzlar experience had gone into the making of this novel: Herder's scathing comments on his young pupil's lack of formal- and self-mastery; the recent indictment by G.E. Lessing of the Neoplatonic doctrine of artistic creation in *Emilia Galotti;* a passing attraction to Maximiliane, the daughter of the German novelist Sophie von La Roche, who probably endowed his heroine with her black eyes. And it was only when Kestner reported the suicide of a Wetzlar acquaintance who had killed himself out of hopeless love that all this was precipitated into a plot. If *Werther* took the world by storm it was because, in Thomas Carlyle's words, it gave expression to "the nameless unrest and longing discontent which was then agitating every bosom." But this first novel is no sentimental tearjerker. Nor is disappointed love its real theme. It is rather what the 18th century called Enthusiasm: the fatal effects of a predilection for absolutes, whether in love, art, society, or the realm of thought. The mind that conceived its symmetry, wove its intricate linguistic patterns, and handled the subtle differentiation of hero and narrator was moved by a formal as well as a personal passion. Even the title has been trivialized in translation: *Sorrows* (instead of "Sufferings") obscures the allusion to the Passion of Christ and individualizes what Goethe himself thought of as a "general confession," in a tradition going back to St. Augustine.

Lyrical and dramatic hymns

Besides *Werther* and *Götz,* the period 1771–75 saw the appearance of a number of magnificent hymns—lyrical or dramatic, according to whether the influence of Pindar or Shakespeare prevailed—"Cäsar," "Mahomets Gesang" ("Mahomet's Singing"), "Der Ewige Jude" ("The Eternal Jew"), "Prometheus," "Sokrates," "Satyros," "Der Wandrer" ("The Wanderer"); the inception of *Egmont* and *Faust* (this so-called *Urfaust,* or "original" version of *Faust,* was discovered by a lucky chance in 1887); the completion of *Clavigo,* a play of more "regular" form on a theme of the French playwright Beaumarchais, and of *Stella* (1775), with its conciliatory ending of a *mariage à trois,* subsequently conventionalized into tragedy. Two operettas, *Erwin und Elmire* and *Claudine von Villa Bella,* reflect a return to the elegance of Rococo inspired by Goethe's betrothal to Lili Schönemann, daughter of a rich banker, who moved in fashionable circles that were soon to prove unbearably restrictive to the young Stürmer und Dränger. From the conflicts of this love he took refuge, as so often, in nature; and in a poem written on the lake of Zürich, "Auf dem See" ("On the Lake"), created the first of those many short lyrics in which language of radiant simplicity is made the vehicle of inexhaustible significance.

With his departure for Weimar in November 1775, the engagement was allowed to lapse.

The mature years at Weimar. Going to Weimar was the major turning point of Goethe's life. He went on a visit to the reigning duke, Charles Augustus. It remained his home—despite Napoleon's invitation to Paris—until his death there on March 22, 1832. From now on, mastery of life became his chief concern; and *Wilhelm Meisters Lehrjahre* (*Wilhelm Meister's Apprenticeship;* 1824), the title he eventually gave his next novel (1795–96), suggests the long apprenticeship such mastery involves. He served his own in the innumerable and ever increasing official duties the young duke heaped on his willing shoulders until, as indispensable minister of the little state, he was inspecting mines, superintending irrigation schemes, and even organizing the issue of uniforms to its tiny army.

Friendship with Charlotte von Stein

He served his apprenticeship, too, in his passionate devotion to the wife of a court official, Charlotte von Stein. For the first time he found himself in love with a woman who could also meet him on the intellectual plane. From the 1,500 or so letters he wrote her we can see her become the guiding principle of his life, teaching him the graces of society, dominating the details of his daily existence, engaging his imagination and desire, yet insisting on a relation governed by decorum and conventional virtue. She would be his sister and nothing more, and the sublimation she increasingly enforced on him, though irksome, could inspire the almost psychoanalytical probings of "Warum gabst du uns die tiefen Blicke?" ("Why did you give us the deep glances?"), the tortures of Orestes and their assuagement by Iphigenie, the delicate one-act play, *Die Geschwister* ("Brother and Sister"; 1776), and such well-loved lyrics as "An den Mond" ("To the Moon"), "Der Becher" ("The Cup"), "Jägers Abendlied" ("Hunter's Evening Song"), "Seefahrt" ("Sea Journey"), and the two exquisite "Wandrers Nachtlieder" ("Wanderer's Night Songs").

In these and other poems of this period—"Grenzen der Menschheit" ("Limits of Mankind"), "Gesang der Geister über den Wassern" ("Singing of the Spirits over the Water"), "Das Göttliche" ("The Divine"), "Harzreise im Winter" ("Journey in the Harz Mountains in Winter"), "Ilmenau"—nature has ceased to be a mere reflection of man's moods and has become something existing in its own right, a setting for an idea or a force indifferent, even hostile to him. This new "objectivity" is in tune with Goethe's growing scientific preoccupations. Yet such is his versatility that he could, when he chose, revert to the temper of "Der König in Thule" ("The King in Thule"; written in 1774) and compose ballads such as "Erlkönig" ("King of the Elves") or "Der Fischer" ("The Fisherman"), in which nature bears the projection of unconscious forces; while a number of *Singspiele,* or musical plays, betoken his readiness and ability to provide light entertainment for the court. *Der Triumph der Empfindsamkeit* ("The Triumph of Sensibility") even satirizes the sensibility his own *Werther* had helped to foster.

But neither the cares of state nor those of a frustrating love affair were conducive to the peace and leisure required to complete works of such magnitude as *Egmont, Faust, Tasso,* and *Iphigenie* (a prose version of this last was sufficiently advanced to be put on before the court in 1779 with Goethe himself in the role of Orestes). And in September 1786, in dramatic secrecy and with the haste of one pursued, he set out on his long-postponed Italian journey. This flight was at once a death and a rebirth. And it was in these terms that he wrote of it in his letters. He sought the renewal of himself, both as man and artist, and so deliberately cut himself off from his emotional, literary, and cultural past, scorning the "Gothic follies" he had once acclaimed, rejecting Juliet's tomb in Verona in favour of the Greek steles in the museum, finding delight in Palladio's churches rather than in San Marco or the doge's palace, devoting barely three hours to Florence, and ignoring completely the medieval glories of Assisi for the sake of its temple of Minerva, feverishly bent on arriving in Rome, "capital of the ancient world," but seeing even that as a prelude to Magna Graecia, to the temples of Paestum, and the revelation of classical grandeur in Sicily, "key to the whole," a prelude to the world of Homer, which he recaptured in a glorious dramatic fragment, *Nausikaa* (1787). And just as he sought and found the *Urmensch,* or archetypal man, in the forms of Greek antiquity, so in these landscapes there came to his mind the extension of this idea to plants as well. In his literary work these pursuits led to the creation of beings who are individual manifestations but of a clearly discernible type; to themes that are universal and timeless but treated in a highly differentiated way; to the measured cadences of verse that are yet vibrant with personal passion.

Italian influence

This new conception of form is apparent in the revision of the four plays he had taken with him to Italy. *Faust, Ein Fragment* ("Faust, a Fragment"), published in 1790, is quite clearly, by its excisions as well as its additions, a step in the direction of the stupendous cultural symbol the play would eventually become rather than any attempt to weld into dramatic unity the sharply individualized episodes of the original version, the *Urfaust. Egmont,* though not actually cast into verse, is raised to the level of poetic drama not by virtue of its frequent iambic rhythms but by a thickening of the verbal texture, so that when music finally takes over it seems the inevitable culmination of a gradual convergence and sudden contraction of themes rather than the "*salto mortale* (*i.e.,* somersault) into the world of opera" Schiller was to dub it. By such means, the personal and the political aspects of the problem become completely interfused—Egmont and his beloved Klärchen, the most lovable characters Goethe ever created, are embodiments of an inner freedom that is a heightened form of the easygoing independence of the Netherlands people—and what had started as a dramatic portrayal of a daemonic individual is transformed into a tragedy of the very idea of freedom, of its fate in a world ruled not just by calculation or intrigue but by unpredictable conjunctures of persons and events.

Torquato Tasso

In *Torquato Tasso* such linguistic density is carried to lengths possible only in verse. Goethe spoke of having expended a positively "unlawful care" on it. But this is not inappropriate to a play about a poet, an artist whose medium is the ordinary vehicle of communication between men. The tragic conflict here arises from misunderstandings about the various modes of language, and the temperamental clashes are presented as concomitants of this rather than as the prime focus of interest (though there is enough psychology to justify the description by the French writer Mme de Staël of Goethe as "*le Racine de l'Allemagne*"). The slightness of the outward action in *Torquato Tasso* has been much criticized, but it can be justified in a study of the "poetical character" per se—a creature for whom "any little vexation grows in five minutes into a theme for Sophocles." By placing him in a society that, far from being indifferent or hostile, cherishes him and values his work, Goethe has thrown into sharpest relief the incurable "discrepancy" between poet and world, and this rift is not healed by Tasso's discovery that even the extremes of anguish can be transmuted into imperishable verse.

Iphigenie

But it was perhaps *Iphigenie auf Tauris* (1787) that benefitted most from his encounter with classical antiquity. And yet Schiller was right in calling it "astonishingly modern and un-Greek." Like *Tasso,* it too treats of the problems of communication: of the unforeseeable power of words once they are released into the world; of the double face of language, which conceals as much as it reveals; of truth, whose opposite is not just an outright lie but the withholding of self. But it treats, too, of man's power to free himself from his myths by recognizing them as projections of his own unconscious, of his power to break the chain of events that seems to determine his present (symbolized in the monotonously regular crime sequence of the race of Tantalus) by a reorientation of outlook. The conciliatory ending, which Euripides contrived by the sudden appearance of the goddess Athena, here comes with the apparent suddenness of new insight: the words of the oracle are susceptible to a different interpretation. In its synthesis of Greek and Christian values, its elevation of the physical to the spiritual through the identification of Iphigenie with the divine sister, Diana, this play represents the highest achievement of 18th-century humanism.

Römische Elegien

The chief lyrical product of the Italian journey was the *Römische Elegien* ("Roman Elegies"; written 1788–89). In their plastic beauty and unabashed sensuality, their blending of erotic tenderness with an enhanced sense of our cultural heritage, these pagan, highly civilized poems are unique in any modern language. Had they been written in the metre of Byron's *Don Juan,* Goethe acknowledged, they might easily have been offensive; but the classical distichs (couplets) lend them that veil of aesthetic distance that reveals even as it shrouds. The true begetter of these elegies was not some passing Roman amour but Christiane Vulpius, daughter of a humble official, whom Goethe had taken into heart and home soon after his return from Italy in April 1788. Christiane bore him several children; but it was not until 1806, when life and property were threatened by the French invasion, that the nonconformist eventually conformed and in grateful recognition of its indissoluble bonds regularized their union in the eyes of society.

His first Italian journey finally brought home to Goethe that, for all his interest and talent, he was not destined to be a painter. Despite diligent practice with his artist friends in Rome, he was never able to master this medium to the point at which it became expressive of his deepest feeling, and with rare exceptions his numerous drawings have no more than the charm of a sensitive amateur. But his abiding preoccupation with the visual arts left an indelible mark on his literary as well as his scientific work and gave added precision to his many critical and aesthetic essays. And it was on this first visit to Italy, too, that he finally reached the decision that he must shed his administrative duties and devote himself henceforth to his true vocation of literature and science.

Political writings during the French Revolution

A return visit to Italy in 1790 brought nothing but disappointment, and a restlessness aggravated by the revolutionary events in the outer world. The *Epigramme. Venedig 1790.* ("Venetian Epigrams of 1790") reflect something of this discontent. In 1792 Goethe accompanied his duke on the disastrous campaign into France, was present at the battle of Valmy, and wrote up his experiences in two still very readable war books, *Campagne in Frankreich 1792* and *Belagerung von Mainz* ("Siege of Mainz"). His liberal-conservative attitudes found expression in *Reineke Fuchs* ("Reynard the Fox"), a recasting of the Low German satire, the *Unterhaltungen deutscher Ausgewanderten* ("Conversations of German Emigrants"), and three plays. *Der Gross-Cophta, Die Aufgeregten* ("The Agitated"), and *Der Bürgergeneral* ("The Citizen General"), which, though artistically unsuccessful, are of interest in being among the few examples of political literature produced by German poets. But it was only as the French Revolution receded that he was able to transmute its overwhelming actuality into timeless poetry. It still forms the background of his Homeric treatment of the refugee problem, *Hermann und Dorothea* (1797). It fills the whole canvas of *Die Natürliche Tochter* ("The Natural Daughter"; 1804). Planned as a trilogy but never completed, this was Goethe's final reckoning with the greatest event of his time. Beneath the coolness of its formal perfection there stirs a profound concern with revolutionary phenomena, with the role of death and destruction in the perpetuation of social and cultural, no less than of natural, forms of life.

Schiller and the classical ideal. The human and spiritual isolation in which Goethe found himself on his return from Italy was unexpectedly relieved by the development of a friendship with Schiller. His acceptance of a formal invitation to contribute to a new journal, *Die Horen* (1795–97; "The Horae"), called forth Schiller's now-famous letter of August 23, 1794, in which, with marvelous insight, he summed up Goethe's whole existence. Here, it seemed to him, was the very embodiment of the naive poet—but consciously naive, moving from feeling to reflection and then transforming reflection back into feeling, concepts of the mind back into percepts of the senses. It was this conscious assent to a mode of thinking different from Schiller's own more abstractive reflection that made possible their immensely fruitful partnership, and the four volumes of their daily correspondence offer not only an invaluable commentary on the ideals and achievements of the greatest period of German literature but astonishing insight into the processes of artistic creation. Some of the works Goethe produced during the next few years are embodiments of their classical ideal. *Hermann und Dorothea,* one of the best loved, is his attempt to "produce a Greece from within." In it he claimed to have "separated the purely human from the dross." The characters are types—except for the hero and heroine, they have no proper names, and even theirs are symbolic—and like those of the *Odyssey* they vindicate peace and home and the domestic virtues. Yet, as always in Goethe's works, these are shown as never secure for long, as constantly in need of being fostered by man's efforts to be human and humane. In the Helena act of *Faust, Part II,* in which the meeting and mating of Faust and Helen of Troy marks the synthesis of paganism and Christianity, of Greece and Germany, he captured the Greek spirit so successfully that competent critics hold that if translated into Attic Greek it might well pass for a lost fragment of the Athenian stage.

A never completed epic, *Achilleis,* is his last attempt to "be a Greek after his own fashion." Other works of this period are in tune with Schiller's growing conviction that the only future for literature in a world that increasingly clamoured for the naturalistic and the tendentious lay in a hermetic closing of the poetic world by a frank introduction of symbolic devices. *Wilhelm Meisters Theatralische Sendung* ("Wilhelm Meister's Theatrical Mission"; a manuscript of this version turned up in 1910) is now widened to a vocation for life, a theme dear to the heart of Schiller, who had himself just completed a treatise *Über die ästhetische Erziehung des Menschen in einer Reihe von Briefen* (1795; "On the Aesthetic Education of Man in a Series of Letters") and wholly in tune with their joint conviction that art, though not the handmaid of either truth or morality, has nevertheless its own peculiar part to play in making better men and better citizens. Fictional realism is now blended with abstraction; characterization, however psychologically acute, subordinated to an overall poetic significance; and the presence in a novel of contemporary society of such mysteriously compelling figures as the Harper and Mignon seems to justify Goethe's claim that his novel is "thoroughly symbolic."

Schiller's influence on *Faust*

It was Schiller, too, who turned his thoughts to the continuation of *Faust* and discerned the difficulties involved in reconciling this "barbarous composition" with their classical ideal, in blending the evident seriousness of its "idea" with that element of "play" that was the prerequisite of the art of the future. By his insistence on such problems, he inspired the fictional framework of *Faust's* "Prelude on the Stage" no less than the philosophical framework of the "Prologue in Heaven." If, in spite of such indications, the world insisted on reading *Faust, Part I* (1808) as a love story, which stamped its author as a Romantic, it was because at this stage the almost unbearable pathos of the Gretchen tragedy had not yet found its place in the wider tragedy of Western man.

Goethe and Schiller blamed the failure of the journals in which they strove to propagate their ideals of art and literature (Goethe's *Propyläen,* 1798–1800, was a quasi-successor to Schiller's *Horen*) on the indifference of an uncultivated public and vented their disappointment in *Xenien,* approximately 400 mordant distichs in the manner of Martial. A more positive reply to their detractors was a wonderful harvest of ballads. Goethe's own—"Der Schatzgräber" ("The Treasure Digger"), "Die Braut von Korinth" ("The Bride from Corinth"), "Der Zauberlehrling" ("The Sorcerer's Apprentice")—differ from his earlier ones in that man rather than nature now holds sway. The "white" magic of reflection is consciously, even ironically, introduced. And in the ballad, with its blend of lyric, epic, and dramatic elements, Goethe now discerned the *Urei,* or archetypal form, of poetry by analogy with the *Urpflanze* (archetypal plants) he had discovered in the vegetable world.

Goethe's relation to the Romantics. With Schiller's death in 1805, Goethe felt he had lost "the half of his existence," and he wrote a magnificent tribute to his great friend in *Epilog zu Schillers Glocke* ("Epilogue to Schiller's Bells"). His intellectual loneliness was eased in some measure by his relations to the new school of Romantics

then flourishing in Jena, for they had much in common. Friedrich von Schlegel had begun his career with a book extolling Greek culture and gone on to praise the Orient as the summit of Romantic thought and poetry. His brother Wilhelm's absorption in form and metre was after Goethe's own heart, and he could not be indifferent to their enthusiastic praise of *Wilhelm Meister* or to Novalis' description of him as "the viceregent of poetry upon earth." In Bettina Brentano, daughter of his old love, Maximiliane von La Roche, he found an ardent response to both his genius and his humanity, and her *Briefwechsel Goethes mit einem Kinde* (1835; "Goethe's Correspondence with a Child") remains one of the most readable books in German literature, whatever doubts may be cast on its reliability. Though Goethe decried the Romantics as "forced talents," amateurishly oblivious of the virtues of form, though he deplored their catholicizing tendencies, their uncritical addiction to all things medieval, their attempts to blur the literary genres and confuse the boundaries between art and life, he yet remained open to many of their enthusiasms, even letting himself be moved to a renewed interest in Gothic architecture. And in *Die Wahlverwandtschaften* (1809) he drew heavily for his thematic material upon their preoccupation with "the night-side of nature," with the animal, magnetic affinities that attract human beings to each other, as elements are attracted in the chemical world.

Die Wahlverwandtschaften

But this novel offers no support at all for a superstitious surrender to forces natural or supernatural, for a subhuman abdication of moral responsibility. Catastrophe follows inexorably upon the arbitrary interpretation of signs and portents; the heroine enters upon a path of renunciation that brings her near sainthood; marriage may be presented with ruthless realism as "a synthesis of impossibilities," but it remains nevertheless "the beginning and end of all civilization." The Romantics were here taught a lesson of social behaviour—and of artistic form. The narrative is conducted with a serene impartiality, and all the classical values of plasticity, restraint, and symmetry are brought to bear on a subject that is sensational to the point of improbability.

By their translations—Romanticism *is* translation, Clemens Brentano declared—the Romantics were opening up the literary treasures of the world, and *Weltliteratur* was to become one of Goethe's most treasured concepts. Its aim was, as he put it, to advance civilization by encouraging mutual understanding and respect—whether through translation or criticism (his own attempts to interpret Serbian poetry to the Germans is an excellent example of this latter) or through the blending of different literary traditions. Two great ballads, "Der Gott und die Bajadere" ("God and the Dancing Girl") and "Paria" ("Outcast"), and two exquisite cycles, the late and lesser known *Chinesisch-Deutsche Jahres- und Tageszeiten* ("Chinese-German Hours and Seasons"; 1830) and the *West-östlicher Divan* (1819), are his own outstanding attempts to marry East with West. This latter is a book of love in all its aspects—tender, playful, sensuous, ironic, wise, and wanton—all of it irradiated by that quality of *Geist*—of intellect, spirit, wit—which he discerned as "the predominant passion" of Persian poetry. His living muse this time, Marianne, the young wife of his friend von Willemer, was perhaps the most completely satisfying of all his loves, so attuned to him in spirit that she could even take a hand in the creation of some of these poems.

The last decade. But the world vision of the aging poet did not only find expression in a silent communing with the past. In his last years, Goethe found himself a world figure, and little Weimar became a Mecca that drew a constant stream of pilgrims from both the Old World and the New. Reports of his stiffness and reserve in the face of almost daily invasions are far outweighed by the testimony of those to whom he showed warmth, understanding, an insatiable curiosity about what was going on in the outside world, and an abiding openness to the present and the future. This is nowhere more apparent than in *Wilhelm Meisters Wanderjahre* (1821–29; "Wilhelm Meister's Travels"), with its commitment to social and technological progress (what he would most like to see before he died, Goethe once said, was the completion of the Panama and Suez canals), to a type of education better adapted to modern specialization than the old humanistic studies, to a world no longer centred wholly in Europe—a major "complication" of his plot is a resettlement plan for emigrants in the land of the future ("Amerika, du hast es besser!" ["America, you are better off!"]). *Wilhelm Meister* points the truth that mastery of life is not conferred at the end of the "apprentice years" and henceforth an inalienable possession, but a ceaseless wandering in which the goal turns out to be the way, and the way the goal.

Wilhelm Meister

At first sight the subtitle, *Die Entsagenden* ("The Renunciants"), seems curiously at odds with such purposeful unrest. But renunciation for Goethe implies no passive resignation to the status quo. It is a growing acceptance of the limits imposed by life itself, limits arising from the nature of space and time and from the conflict of interests and potentialities. The apparent formlessness of the novel reflects the duality of its title. It meanders, its narrative interspersed with tales, anecdotes, episodes and maxims, having but the loosest connection with the plot but a formal, if often subterranean, connection with the poetic significance. These interpolations, like the increasingly symbolic characters, display the whole spectrum of human modes of renunciation. The "whole man" is here represented not by any single individual but by a constellation of many, and the informing principle is the spatial one of configuration rather than the temporal one of succession.

Faust, too, is often decried as formless, though the climate of criticism is now more propitious to the discovery of its "law." The array of lyric, epic, dramatic, operatic, and balletic elements, of almost every known metre, from doggerel through terza rima (an Italian form of iambic verse consisting of stanzas of three lines) to six-foot trimeter (a line of verse consisting of three measures), of styles ranging from Greek tragedy through medieval mystery, baroque allegory, Renaissance masque, commedia dell'arte, and the "temerities of the English stage," to something akin to the modern revue, all suggest a deliberate attempt to make these various forms a vehicle of cultural comment rather than any failure to create a coherent form of his own. And the content with which Goethe invests his forms bears this out. He draws on an immense variety of cultural material—theological, mythological, philosophical, political, economic, scientific, aesthetic, musical, literary—for the more realistic Part I no less than for the more symbolic Part II (first published posthumously in 1832): if Faust's wooing of Helena in the "Classic-Romantic Phantasmagoria" (as the first publication of the scene in 1827 called it) is accomplished by teaching her the unfamiliar delights of rhymed verse, his seduction of Gretchen is firmly set in the long tradition of erotic mysticism going back to the Song of Solomon. The Faust myth is here made the medium of a profoundly serious but highly ironic commentary on our cultural heritage, presented not as historical pageant—Faust's "progress" from his 18th- to 16th-century beginnings back through the Middle Ages and classical antiquity to the origins of life, and beyond that to the "Mothers," timeless source of all forms of being, annuls the historical time sequence—but as a drama of the diverse potentialities that coexist in Western civilization.

Faust

This Faust, unlike his creator, is the very type of Western man, with two souls warring within his breast and a restlessly inquiring spirit. To the 19th century his ceaseless striving seemed a good thing in itself. To a generation shocked into doubts about progress and the value of action, the disastrous consequences of his attempts to experience "the weal and woe of all mankind" (the *libido sciendi* of Marlowe's Faustus is here but briefly indulged and as swiftly transcended) loom larger than the quotable "message" of any of the speeches, and his ultimate "salvation" becomes correspondingly suspect. Yet the love that bears his mortal remains to "higher spheres" does not mitigate the ironic defeat of his highest mortal endeavour. If the seal of approval is set on a spirit that has eluded Mephisto's every effort to lull him into sloth, the evil into which it led him is not condoned. It needs the combined intercession of human wisdom and human suf-

fering, human innocence and human experience, before compassionate verdict is passed on the erring and straying of this soul "in ferment." Indeed, none of Goethe's conciliatory endings, except that of *Iphigenie,* really removes the sting of tragedy. Critics have tended to excuse or deplore them by reference to his own *konziliante Natur* (his "conciliatory nature"). But at least as relevant is his preoccupation with the form of Greek trilogies and tetralogies and his unorthodox interpretation of Aristotle's catharsis as an effect only likely to be produced in the spectator if there is a corresponding element of "reconciliation" in the structure of the play itself. The apotheosis of the hero, whether Faust's, Egmont's, or Ottilie's in the *Wahlverwandtschaften,* is always set in a context reminiscent of a theophany and of the ritual origins of tragedy.

Interest in music

Nor can his interest in the cathartic effect of music be ignored. Unlike the German Romantic poet Novalis, for whom music was "the key to the universe," Goethe was profoundly aware of its dual nature and as suspicious as Plato of its orgiastic power. As in every art he looked for the taming of the Dionysiac by the Apolline, nowhere more movingly symbolized than by the taming of the lion through the piping of the little child in his *Novelle* of 1828, a theme he had already discussed with Schiller as far back as 1797. And increasingly he turned to music for assuagement of his own suffering. His *Trilogie der Leidenschaft* ("Trilogy of Passion"; 1823–27) is at once the lyrical precipitate of an old man's anguished love for a girl of 18 and a tribute to the cathartic effect of this "heavenly art," which restores to life even as it soothes. His *Zauberflöte, Zweiter Teil* is a tribute to his favourite Mozart's *Magic Flute:* Mozart would, he thought, have been the ideal composer for *Faust.* And one of the comforts of his later years was an intimate friendship with the composer K.F. Zelter, whose most brilliant pupil, the young Mendelssohn, afforded him hours of musical delight and deepened his musical understanding—though he never succeeded in reconciling him to the daemonic aspects of Beethoven's music.

Scientific work

By common consent, *Faust* is one of the supreme, if as yet unclassified, achievements of literature. But there were moments when Goethe rated his scientific work higher than all his poetry. His predilection for his *Farbenlehre* ("Theory of Colour"; 1805–10) has something of the love of a parent for a problem child, and nothing is easier than for the physicist to pick holes in his systematic attempt to prove Newton wrong, or for the psychologist to find the cause of his stubbornness in his sense of mathematical inadequacy or in his neurotic attachment to the doctrine that light is one and indivisible and never to be explained by any theory of particles. On the other hand, the usefulness of the Psycho-Physiological Section, together with his study *Entoptische Farben* ("Entoptic Images"), is generally acknowledged, while the Historical Section is something of a pioneer work in the writing of the history of science. His work in botany and biology is less controversial. His *Metamorphose der Pflanzen* ("Attempt to Explain the Metamorphosis of Plants"; 1790) is a model of presentation, and the drawings in it are a botanist's delight. His main thesis, that all the parts of the plant are modifications of a type-leaf, has met with a measure of acceptance, though his categorical neglect of the root is regarded as an unscientific exclusion of a possible area of relevance. His hypothesis of a type-plant, by contrast, commands no interest among orthodox botanists today. His discovery in 1784, arrived at independently even if he was not the first to make it, of a recognizable os intermaxillare (the premaxilla of modern anatomists) in the human species was yet another result of his sustained quest for unity and continuity in nature and caused Darwin to hail him as a forerunner.

But what makes for the continuing interest of Goethe's science is not his discoveries: he could not always claim priority for them at the time, nor was he in the least interested in doing so. It is his insight into his methods of arriving at them. Few have been as aware of the mental processes involved in the study of natural phenomena; few have been more alive to the hazards that beset the scientist, at every level, from sheer observation to the construction of a theory; and few have been more conscious of the unwitting theorizing involved in even the simplest act of perception. And no one has argued more convincingly that the only way of coping with this inescapable involvement of the observer in the phenomena to be observed is to let "knowledge of self" develop with "knowledge of world."

Such scrupulous awareness of his own mental operations was, of course, of paramount importance in morphology, the science Goethe founded and named. Morphology, as he understood it, was the systematic study of formation and transformation—whether of rocks, clouds, colours, plants, animals, or the cultural phenomena of human society—as these present themselves to sentient experience. He did not propose it as a substitute for the quantitative sciences, which break down forms as we know them and by converting them into mathematical terms ensure a measure of prediction and control. He was not, contrary to common belief, opposed to analysis—one of his favourite maxims was that analysis and synthesis must alternate as naturally as breathing in and breathing out—and his only objection to physics was its increasing tendency to claim monopoly of understanding. What he was aiming at was rather a humanizing supplement, an understanding of nature in all its qualitative manifestations; and one of his most impassioned pleas is for a concert of all the sciences, a cooperation of all types of method and mind.

This impulse, to find a scientific as well as an aesthetic corrective to the inevitably esoteric tendencies of specialization, is nowhere more apparent than in his two elegies on plant and animal metamorphosis in which he tries to present to imagination and feeling what has been understood by the mind. They eventually took their place in a cycle of philosophical poems entitled *Gott und Welt* ("God and World"). Though no orthodox believer, Goethe was by no means the pure pagan the 19th-century critics liked to imagine. Spinoza's pantheism certainly struck a sympathetic chord, for the Deist idea of a God who, having created the world, then left it to revolve, was repugnant to him. But he was and remained a grateful heir of the Christian tradition—*bibelfest,* rooted in the Bible—as his language constantly proclaims. And it was from this centre that he extended sympathetic understanding to all other religions, seeking their common ground without destroying their individual excellences, seeing them as different manifestations of an *Ur,* or archetypal, religion and thus giving expression, in this field as elsewhere, to the essentially morphological temper of his mind. "Panentheism" has been proposed as a more exact term for his belief in a divinity at once immanent and transcendent, and he rebuked those who tried to confine him to one mode of thought by saying that as poet he was polytheist, as scientist pantheist, and that when, as a moral being, he had need of a personal God, "that too had been taken care of." This was one of the meanings he attached to the biblical text: "In my father's house are many mansions."

Religious views

Appraisal. A day will come, Carlyle predicted in a letter to Ralph Waldo Emerson, when "you will find that this sunny-looking courtly Goethe held veiled in him a Prophetic sorrow deep as Dante's." And since World War II there have been many attempts to replace the image of the serene optimist by that of the tortured skeptic. The one is as inadequate as the other—as inadequate as T.S. Eliot's conclusion that he was sage rather than poet—though this is perhaps inevitable when a writer is such a master of his own medium that even his prose proves resistant to translation. Even his Werther knew that the realities of existence are rarely to be grasped by Either-Or. And the reality of Goethe himself certainly eludes any such attempt. If he was a skeptic, and he often was, he was a hopeful skeptic. He looked deep into the abyss, but he deliberately emphasized life and light. He lived life to the full at every level, but never to the detriment of the civilized virtues. He remained closely in touch with the richness of his unconscious mind, but he shed on it the light of reflection without destroying the spontaneity of its processes. He was, as befits a son of the Enlightenment, wholly committed to the adventure of science; but he stood in awe and reverence before the mystery of the universe. Goethe nowhere formulated a system of thought. He was

as impatient of the sterilities of logic chopping as of the inflations of metaphysics, though he acknowledged his indebtedness to many philosophers, including Kant. But here again he was not to be confined. Truth for him lay not in compromise but in the embracing of opposites. And this is expressed in the form of his *Maximen* ("maxims"), which, together with his *Gespräche* ("conversations"), contain the sum of his wisdom. As with proverbs, one can always find among them a twin that expresses the complementary opposite. And they have something of the banality of proverbs too. But it is, as André Gide observed, "*une banalité supérieure.*" What makes it "superior" is that the thought has been felt and lived and that the formulation betrays this. And for all his specialized talents, there was a kind of "superior banality" about Goethe's life. If he himself felt it was "symbolic" and worth presenting as such in a series of autobiographical writings, it was not from arrogance but from a realization that he was an extraordinarily ordinary man in whom ordinary men might see themselves reflected. Not an ascetic, a mystic, a saint, or a recluse, not a Don Juan or a poet's poet but one who to the best of his ability had tried to achieve the highest form of *l'homme moyen sensuel*—which is perhaps what Napoleon sensed when after their meeting in Erfurt he uttered his famous "*Voilà un homme!*" (E.M.Wn.)

MAJOR WORKS

VERSE: *Das Buch Annette* (1767); *Neue Lieder,* so-called *Leipziger Liederbuch* (1770); *Römische Elegien* (1795); *Venezianische Epigramme* (1790); *Xenien,* written with Schiller (1796); *Balladenjahr* (1797); *Epilog zu Schillers Glocke* (1805); *West-östlicher Divan* (1819); *Trilogie der Leidenschaft* (1823–27).

DRAMATIC WORKS: *Die Laune des Verliebten* (1767, printed 1806); *Die Mitschuldigen* (1768, printed 1787); *Götz von Berlichingen* (1773); *Götter, Helden und Wieland* (1774); *Clavigo* (1774); *Erwin und Elmire* (1775), operetta; *Stella* (1776); *Claudine von Villa Bella* (1776), operetta; *Die Geschwister* (1776); *Der Triumph der Empfindsamkeit* (1777–78); *Iphigenie auf Tauris* (1779; final version, 1787); *Egmont* (1788); *Torquato Tasso* (1790); *Faust, Ein Fragment* (1790); *Der Gross-Cophta* (1792); *Der Bürgergeneral* (1793); *Die Natürliche Tochter* (1804); *Faust I* (1808); *Pandoras Wiederkunft* (1808); *Des Epimenides Erwachen* (1815); *Faust II* (1832).

NOVELS AND EPICS: *Die Leiden des jungen Werthers* (1774); *Reineke Fuchs* (1794), satire; *Unterhaltungen deutscher Ausgewanderten* (1795); *Wilhelm Meisters Lehrjahre* (1795–96); *Hermann und Dorothea* (1797); *Achilleis* (1808), epic fragment; *Die Wahlverwandtschaften* (1809); *Wilhelm Meisters Wanderjahre* (1821–29); *Novelle* (1826).

AUTOBIOGRAPHY: *Aus meinem Leben, Dichtung und Wahrheit* (1811–22); *Italienische Reise* (1816–17); *Campagne in Frankreich, 1792* (1822); *Die Belagerung von Mainz 1793* (1822).

SCIENCE: *Versuch die Metamorphose der Pflanzen zu erklären* (1790); *Beiträge zur Optik* (1791–92); *Zur Farbenlehre* (1810); *Zur Naturwissenschaft überhaupt, besonders zur Morphologie* (1817–23), journal.

CRITICISM: *Von deutscher Baukunst* (1773); *Die Propyläen* (1798–1800), journal ed. by Goethe; *Winkelmann und sein Jahrhundert* (1805); *Über Kunst und Altertum,* 6 vol. (1816–32), periodical.

TRANSLATIONS: There is still no uniform edition in English though the chief works have been frequently translated and many of them are to be found in Bohn's Standard Library, 14 vol. (1846–90). For details see Lina and Eugen Oswald, and Alexander Dickson, *Goethe in England (and America)* in *Publications of the English Goethe Society,* 2 vol. (1909, 1951). Translations include: *Wilhelm Meister's Theatrical Mission* by G.A. Page (1913); *Wilhelm Meister* by Thomas Carlyle (1824; Everyman ed., 1912, reissued 1965). *Faust I and II* by A.G. Latham (1902–06; Everyman ed., 1908); other translations of *Faust* by Bayard Taylor (1870–71; World's Classics, 1932); Philip Wayne (Penguin ed., 1949); Louis MacNeice (1951); Bertram Jessup (1958); John Shawcross (1959); Walter Kaufmann (1961); Alice Raphael (1963); and C.E. Passage (1965). *Egmont* by Michael Hamburger in *The Classic Theatre,* vol. 2, ed. by E. Bentley (1959); by Willard R. Trask (1960); *Götz von Berlichingen,* by C.E. Passage (1965); *Iphigenia in Tauris,* by C.E. Passage (1963); *Italian Journey, 1786–1788,* by W.H. Auden and Elizabeth Mayer (1962); *Torquato Tasso,* by C.E. Passage (1966); *Kindred by Choice (Die Wahlverwandtschaften)* by H.M. Waidson (1961); by Elizabeth Mayer and Louise Bogan (1963); *Werther* by William Rose (1929); *Werther* by B.Q. Morgan (1957); *Die Leiden des jungen Werther,* Dual-Language book, ed. by H. Steinhauer (1962); *Werther: The New Melusina. Novelle* by Victor Lange (1949); *Novelle* by C.P. Middleton (1959); *The West-Eastern Divan* by Edward Dowden (1914). *The Briefwechsel* with Schiller and a collection of early and miscellaneous letters are in Bohn (see above); that with Carlyle, ed. by C.E. Norton (1887). *Letters from Goethe,* selected and trans. by M. von Herzfeld and C.M. Sym (1957). *Conversations with Eckermann* by John Oxenford, 2 vol. (1850; vol. 6 of Bohn); by R.O. Moon (1951).

BIBLIOGRAPHY

Works, drawings, diaries, correspondence, and conversations: An early complete edition is JOHANN WOLFGANG VON GOETHE, *Goethe's Werke: Vollständige Ausgabe letzter Hand,* 55 vol. in 19 (1828–33). The standard critical edition is JOHANN WOLFGANG VON GOETHE, *Goethes Werke: herausgegeben im Auftrage des Grossherzogin Sophie von Sachsen,* 133 vol., including the scientific works, diaries, and letters (1887–1919, reprinted 1975); it was published in Weimar at first under the auspices of the grand duchess Sophie, and hence is known as the *Weimarer Ausgabe* or *Sophienausgabe.* In 1952 the Deutsche Akademie der Wissenschaften zu Berlin undertook a new critical edition, publishing separate works or small collections of works with detailed historical commentary. Of modern editions the most noteworthy are: *Goethes sämtliche Werke: Jubiläums-Ausgabe,* ed. by EDUARD VON DER HELLEN, 40 vol. (1902–07), supplemented by the *Register* (1912), an index volume, with introductions and notes; the *Gedenkausgabe der Werke, Briefe und Gespräche, 28. August 1949,* ed. by ERNST R. BEUTLER, 24 vol., including selections from the scientific works, correspondence, and conversations, with useful indexes (1948–54), and a supplementary volume, *Ergänzungsband I: Briefe aus dem Elternhaus* (1960); and *Werke: Hamburger Ausgabe,* ed. by ERICH TRUNZ, 14 vol. (1961–64), issued with varying edition statements. In 1974 the 10th rev. ed. was begun. There are innumerable collections of selected works, such as *Goethes amtliche Schriften,* ed. by WILLY FLACH, 3 vol. (1950), including Goethe's works of 1776–1819, with commentaries; a later comprehensive selection is JOHANN WOLFGANG VON GOETHE, *Werke,* ed. by EMIL STEIGER, 4 vol. (1982). Goethe's drawings and sketches were published in the *Corpus der Goethezeichnungen,* ed. by GERHARD FEMMEL, 6 vol. (1958–70); see also LUDWIG MUNZ, *Goethes Zeichnungen und Radierungen* (1949); *Goethes Gespräche,* ed. by FLODOARD W. VON BIEDERMANN, 1st ed., 10 vol. (1889–96), and 2nd ed., 5 vol. (1909–11); and JOHANN P. ECKERMANN, *Gespräche mit Goethe in den letzten Jahren seines Lebens,* 3 vol. (1836–48, many times reprinted in 1 vol.). A new edition of conversations, based on Biedermann's collection, was begun in 1965 by WOLFGANG HERWID (ed.), *Goethes Gespräche: eine Sammlung zeitgenössischer Berichte aus seinem Umgang.* The only complete collection of Goethe's letters is in the *Weimarer Ausgabe;* see also the letters to Goethe, ed. by RUDOLF K. GOLDSCHMIDT-JENTNER, *Eine Welt schreibt an Goethe* (1937). His correspondence with particular individuals, German and foreign, is conveniently listed in the *Hamburger Ausgabe,* vol. 14; see also DOUGLAS F.S. SCOTT (ed.), *Some English Correspondents of Goethe* (1949), which includes correspondence with Mathew ("Monk") Lewis, T. Holcroft, Sir Walter Scott, P.P. Gillies, Sir John Bowring, Lord Leveson-Gower, and Sarah Austin. Selections in English include JOHANN WOLFGANG VON GOETHE, *Goethe: Conversations and Encounters,* ed. and trans. by DAVID LUKE and ROBERT PICK (1966), based on the last three volumes of the centenary edition *Gedenkausgabe;* JOHANN WOLFGANG VON GOETHE, *Goethe's World View, Presented in His Reflections and Maxims* (1963), a bilingual text, ed. and with an introduction by FREDERICK UNGAR; JOHANN WOLFGANG VON GOETHE, *The Autobiography of Johann Wolfgang von Goethe,* 2 vol. (1974), trans. by JOHN OXENFORD; JOHANN WOLFGANG VON GOETHE, *Goethe's Literary Essays* (1964), a selection arranged by J.E. SPINGARN; JOHANN WOLFGANG VON GOETHE, *Roman Elegies and Venetian Epigrams* (1974), a bilingual text, trans. and with an introduction and commentaries by LEVI R. LIND; and JOHANN WOLFGANG VON GOETHE, *The Eternal Feminine: Selected Poems* (1980), a bilingual text, ed. by FREDERICK UNGAR.

Bibliography: Only the researcher will consult the vast compilation in KARL GOEDEKE, *Grundriss zur Geschichte der deutschen Dichtung aus den Quellen,* 5 parts (1910–13 and later editions); superseded by HANS PYRITZ, *Goethe-Bibliographie,* new ed., 2 vol. (1965–68). Volume 14 of the *Hamburger Ausgabe* is adequate for most purposes. The *Goethe-Handbuch,* ed. by ALFRED ZASTRAU (publication begun in 1955), provides alphabetical references to all topics relating to Goethe. HEINZ KINDERMANN, *Das Goethebild des XX. Jahrhunderts,* 2nd rev. ed. (1966), is in effect a lively catalogue raisonné of European and American critical literature during the 20th century. Later bibliographies have appeared in numerous periodicals and yearbooks issued by Goethe societies throughout the world, as well as in the annual *MLA International Bibliography of Books and Articles on the Modern Languages and Literatures.*

Goethe societies: Wiener Goethe-Verein, 1886; Goethe

Gesellschaft, Weimar, 1885; English Goethe Society, 1886; Australian Goethe Society, 1949; Goethe Society of Maryland and the District of Columbia, 1932; Japanese Goethe Gesellschaft, 1932. All issue publications.

Biography and criticism: The best introduction in English is GEORGE H. LEWES, *The Life of Goethe,* new ed. (1965). Others available in English are by ALBERT BIELSCHOWSKY, *The Life of Goethe,* 3 vol. (1905–08, reprinted 1970; originally published in German, 2 vol., 1896–1904); BENEDETTO CROCE, *Goethe* (1923, reprinted 1973; trans. from the Italian); and GEORGE M. BRANDES, *Wolfgang Goethe* (1924; originally published in Danish, 1915). See also EMIL LUDWIG, *Goethe: The History of a Man, 1749–1832* (1928, reprinted 1936; originally published in German, 3 vol., 1920); HENRY W. NEVINSON, *Goethe: Man and Poet* (1931, reprinted 1971); JOHN G. ROBERTSON, *The Life and Work of Goethe, 1749–1832,* enlarged ed. (1932, reprinted 1973); LUDWIG LEWISOHN (ed.), *Goethe: The Story of a Man: Being the Life of Johann Wolfgang Goethe as Told in His Own Words and the Words of His Contemporaries,* 2 vol. (1949); ALBERT SCHWEITZER, *Goethe: Five Studies* (1961; trans. from the German); and KARL VIËTOR, *Goethe, The Poet* (1949, reprinted 1970), and *Goethe, the Thinker* (1950), both translations of parts of his *Goethe: Dichtung, Wissenschaft, Weltbild* (1949). Criticism in English includes BARKER FAIRLEY, *Goethe, as Revealed in His Poetry* (1932, reprinted 1973), and *A Study of Goethe* (1947, reprinted 1977); THOMAS MANN, *Three Essays* (1929, reprinted 1932; originally published in German, new ed., 1932); WILLIAM ROSE (ed.), *Essays on Goethe* (1949); ALEXANDER R. HOHLFELD, *Fifty Years with Goethe, 1901–1951: Collected Studies* (1953); ELIZABETH M. WILKINSON and LEONARD A. WILLOUGHBY, *Goethe: Poet and Thinker* (1962); and RONALD PEACOCK, *Goethe's Major Plays* (1959, reprinted 1966). Other studies include RONALD D. GRAY, *Goethe: A Critical Introduction* (1967); VICTOR LANGE (ed.), *Goethe* (1968), a collection of essays; HARRY G. HAILE, *Artist in Chrysalis* (1973), a study of Goethe's Italian period (1786–88); ILSE GRAHAM, *Goethe and Lessing: The Wellsprings of Creation* (1973); LISELOTTE DIECKMANN, *Johann Wolfgang Goethe* (1974), a survey of Goethe's autobiographical works and his studies in the natural sciences; ERNST M. OPPENHEIMER, *Goethe's Poetry for Occasions* (1974); ERIC A. BLACKALL, *Goethe and the Novel* (1976); ERICH HELLER, *The Poet's Self and the Poem* (1976); MEREDITH LEE, *Studies in Goethe's Lyric Cycles* (1978); GEORG LUKÁCS, *Goethe and His Age* (1968, reprinted 1978; originally published in German, 1947); and JÖRN GÖRES, *Goethes Leben in Bilddokumenten* (1981), an illustrated biography.

On Faust: JOHANN WOLFGANG VON GOETHE, *Goethe's Urfaust and Faust,* ed. by LEONARD A. WILLOUGHBY (1943); *Faust,* ed. by ROE-MERRIL S. HEFFNER, HELMUT REHDER, and WILLIAM F. TWADDELL, 2 vol. (1975); and HERMANN G. FIEDLER, *Textual Studies of Goethe's Faust* (1946, reprinted 1977). Further interpretations of Faust in English are by F. MELIAN STAWELL and G. LOWES DICKINSON, *Goethe & Faust* (1928, reprinted 1972); RONALD D. MILLER, *The Meaning of Goethe's Faust* (1939); DENNIS J. ENRIGHT, *Commentary on Goethe's Faust* (1949, reprinted 1977); BARKER FAIRLEY, *Goethe's Faust: Six Essays* (1953, reprinted 1965); ALEXANDER GILLIES, *Goethe's Faust* (1957); and STUART P. ATKINS, *Goethe's Faust: A Literary Analysis* (1958, reprinted 1964). See also ELIZA M. BUTLER, *The Fortunes of Faust* (1952, reprinted 1979); and ADOLF I. FRANTZ, *Half a Hundred Thralls to Faust: A Study Based on the British and the American Translators of Goethe's Faust, 1823–1949* (1949). Other works include JOHANN WOLFGANG VON GOETHE, *Faust,* trans. by BARKER FAIRLEY (1970), a prose translation; EUDO C. MASON, *Goethe's Faust: Its Genesis and Purport* (1967); PETER SALM, *The Poem as Plant: A Biological View of Goethe's Faust* (1971); PIETRO CITATI, *Goethe* (1974; originally published in Italian, 1970), a companion to *Faust* and *Wilhelm Meisters Lehrjahre;* HARRY G. HAILE, *Invitation to Goethe's Faust* (1978); KURT WEINBERG, *The Figure of Faust in Valéry and Goethe* (1976); HAROLD JANTZ, *The Mothers in Faust: The Myth of Time and Creativity* (1969), and *The Form of Faust: The Work of Art and Its Intrinsic Structures* (1978); JOHN GEAREY, *Goethe's Faust: The Making of Part I* (1981); ALLAN P. COTTRELL, *Goethe's Faust* (1976), and *Goethe's View of Evil and the Search for a New Image of Man in Our Time* (1982).

Science, philosophy, and aesthetics: SIR CHARLES SHERRINGTON, *Goethe on Nature and on Science,* 2nd ed. (1949), is an unsympathetic view; as a corrective to Sherrington, see AGNES ARBER, *Goethe's Botany,* including a trans. of *Die Metamorphose der Pflanzen* (1946), and of *The Natural Philosophy of Plant Form* (1950, reprinted 1970). See also RENÉ MICHÉA, *Les Travaux scientifiques de Goethe* (1943); RONALD D. GRAY, *Goethe, the Alchemist* (1952); ERICH HELLER, *The Disinherited Mind,* 4th ed. (1975); OTTO HARNACK, *Goethe in der Epoche seiner Vollendung, 1805–1832,* 3rd ed. (1905); GEORG SIMMEL, *Goethe* (1913, reprinted 1923); LANCELOT L. WHYTE, *The Next Development in Man* (1944, reprinted 1962); CARL J. OBENAUER, *Goethe in seinem Verhältnis zur Religion* (1921); HANS M. WOLFF, *Goethes Weg zur Humanität* (1951); FRITZ-JOACHIM VON RINTELEN, *Der Rang des Geistes: Goethes Weltverständnis* (1955); PAUL STÖCKLEIN, *Wege zum späten Goethe: Dichtung, Gedanke, Zeichnung,* 2nd ed. (1960, reprinted 1977); IJS JOLLES, *Goethes Kunstanschauung* (1957). Selections in English include JOHANN WOLFGANG VON GOETHE, *Wisdom and Experience,* ed. and trans. by HERMANN J. WEIGAND (1949); GÜNTHER MÜLLER (ed.), *Maximen und Reflexionen* (1947); THOMAS MANN (ed.), *The Permanent Goethe* (1948); and WOLFGANG LEPPMANN, *The German Image of Goethe* (1961). Goethe's aesthetics is discussed in JOHANN WOLFGANG VON GOETHE, *Goethe's Theory of Colours,* trans. and with notes by CHARLES LOCK EASTLAKE (1975); DEREK VAN ABBÉ, *Goethe: New Perspectives on a Writer and His Time* (1972); JOHANN WOLFGANG VON GOETHE, *Goethe on Art,* ed. and trans. by JOHN GAGE (1980); GABRIELE GIRSCHNER, *Goethes Tasso: Klassizismus als ästhetische Regression* (1981); WILLIAM D. ROBSON-SCOTT, *The Younger Goethe and the Visual Arts* (1981); TERENCE J. REED, *The Classical Centre: Goethe and Weimar, 1775–1832* (1980); and WALTER KAUFMANN, *Goethe, Kant, and Hegel* (1980).

Weltliteratur: JEAN M. CARRÉ, *Goethe en Angleterre: Étude de littérature comparée* (1920); WILLIAM ROSE, *From Goethe to Byron: The Development of "Weltschmerz" in German Literature* (1924); JOHN G. ROBERTSON (ed.), *Goethe and Byron* (1925, reprinted 1978); STUART P. ATKINS, *The Testament of Werther in Poetry and Drama* (1949); ELIZA M. BUTLER, *Byron and Goethe: Analysis of a Passion* (1956); FREDERICK NORMAN, *Henry Crabb Robinson and Goethe,* 2 vol. (1930–32); JAMES B. ORRICK, *Matthew Arnold and Goethe* (1928, reprinted 1977); FRITZ STRICH, *Goethe and World Literature* (1949, reprinted 1972; originally published in German, 1946, reprinted 1957); HUMPHRY TREVELYAN, *Goethe and the Greeks* (1941, reprinted 1981; a corrective to ELIZA M. BUTLER, *The Tyranny of Greece over Germany: A Study of the Influence Exercised by Greek Art and Poetry over the Great German Writers of the Eighteenth, Nineteenth, and Twentieth Centuries,* 1935, reprinted 1958); JAMES BOYD, *Goethe's Knowledge of English Literature* (1932, reprinted 1973); BERTRAM BARNES, *Goethe's Knowledge of French Literature* (1937); HORST OPPEL, *Das Shakespeare-Bild Goethes* (1949); ARNOLD FEDERMANN, *Der junge Goethe und England* (1949); ANDRÉ VON GRONICKA, *The Russian Image of Goethe: Goethe in Russian Literature of the First Half of the Nineteenth Century* (1968); and DAVID B. RICHARDS, *Goethe's Search for the Muse: Translation and Creativity* (1979).

Cultural background: HERMANN A. KORFF, *Geist der Goethezeit,* 5 vol. (1923–57, reprinted 1974); ERNST R. BEUTLER, *Essays um Goethe,* 7th ed., 2 vol. (1980); *Goethe und siene Welt,* with 580 illustrations, ed. by ERNST R. BEUTLER, HANS WAHL, and ANTON KIPPENBERG (1932); ARNOLD BERGSTRÄSSER, *Goethe's Image of Man and Society* (1949, reprinted 1962); WALTER H. BRUFORD, *Theatre, Drama, and Audience in Goethe's Germany* (1950, reprinted 1974), and *Culture and Society in Classical Weimar, 1775–1806* (1962); JOHN PRUDHOE, *The Theater of Goethe and Schiller* (1973); and MARVIN CARLSON, *Goethe and the Weimar Theatre* (1978).

(E.M.Wn./Ed.)

The Forms of Government: Their Historical Development

Most of the key words commonly used to describe governments, words such as monarchy, oligarchy, and democracy, are of Greek or Roman origin. They have been current for more than 2,000 years and have not yet exhausted their usefulness. This suggests that mankind has not changed very much since they were coined; but such verbal and psychological uniformity must not be allowed to hide the enormous changes in society and politics that have occurred. The earliest analytical use of the term monarchy occurred in ancient Athens, chiefly in Plato's dialogues, but even in Plato's time the word was not self-explanatory. There was a king in Macedon and a king in Persia, but the two societies, and therefore their institutions, were radically different. To give real meaning to the word monarchy in these two instances, it would be necessary to investigate their actual political and historical contexts. Any general account of monarchy required then, and requires today, an inquiry as to what circumstances have predisposed societies to adopt monarchy, and what have led them to reject it. So it is with all political terms.

This article is divided into the following sections:

Primitive government 141
- Agricultural society
- The spread of civilization

Greece 142
- The city-state
- Monarchy, oligarchy, democracy

Rome 142
- The republic
- The empire

The Middle Ages 143
- Dissolution and instability
- Feudalism
- The rise of law and the nation-state

Emergence of the modern world 144
- The failure of absolutism
- Representation and constitutional monarchy
- The American and French republics
- Nationalism and imperialism

20th-century models 146
- The Soviet state
- Liberal democracy

Bibliography 147

PRIMITIVE GOVERNMENT

Agricultural society. So long as humankind were few, there was hardly any government. The division of function between ruler and ruled occurred only, if at all, within the family. The largest social groups, whether tribes or villages, were little more than loose associations of families, in which every elder or family head had an equal voice. Chieftains, if any, had strictly limited powers; some tribes, no doubt, did without chieftains altogether. This prepolitical form of social organization may still be found in undeveloped regions of the world, such as the Amazonian jungle in South America or the Upper Nile Valley in Africa.

Pre-political societies

The rise of agriculture began to change this state of affairs. In the land of Sumer (modern Iraq) the invention of irrigation necessitated grander arrangements. Control of the flow of water down the Tigris and Euphrates rivers had to be coordinated by a central authority, so that downstream fields could be watered as well as those further up. It became necessary also to devise a calendar, so as to know when the spring floods might be expected. As these skills evolved, society evolved with them. In early Sumer, it is reasonable to assume, the heads of the first cities, which were little more than enlarged villages, only gradually assumed the special attributes of monarchy, the rule of one; and the village council only gradually undertook a division of labour, so that some specialized as priests and others as warriors, farmers, or tax gatherers (key figures in every civilized society). As organization grew more complex, so did religion: an elaborate system of worship was necessary to propitiate the quite elaborate family of gods who, it was hoped, would protect the city from attack, from natural disaster, and from any questioning of the political arrangements deemed necessary by the ruler group.

Unfortunately, but inevitably, the young cities of Sumer quarrelled over the distribution of the rivers' water, and their wealth excited the greed of nomads outside the still comparatively small area of civilization. War, perhaps the most potent of all forces of historical change, announced its arrival, and military leadership became at least as important an element of kingship as divine sanction. It was to remain so throughout the long history of monarchy: whenever kings have neglected their military duties they have endangered their thrones. The wars of Sumer also laid bare another imperative of monarchy—the drive for empire, arising from the need to defend and define frontiers by extending them, and the need to find the means to pay for troops and weapons, whether by the plunder of an enemy or by the conquest of new lands.

Military and imperial imperatives

The spread of civilization. The history of Old World monarchy, and indeed of civilization, was to consist largely of variations on these patterns for four or five millennia. Trade contacts carried the principles of civilization to Egypt and to India (China seems to have evolved independently); and everywhere, once the social order was established, the problem of defending it became paramount. For although the broad zone of civilization spread steadily, so that by the reign of the Roman emperor Trajan (AD 98–117) there was a continuous band of civilized societies from Britain to the China Sea, it was always at risk from the barbarian nomads who roamed the great steppelands of central Eurasia. These nomads had retained the loose and simple institutions of humanity's infancy, but they had in other ways evolved as rapidly and successfully as the cities themselves (and partly under the cities' influence). The steppe was horse country, and, armed with bows and arrows, the barbarians of all epochs were marvelously swift and deadly light cavalry. They fought constantly among themselves for pasturage, and the losers were forever being driven west, south, and east, where they all too often overcame such defenses as the farms and cities of civilization could muster against them.

The nomads' military challenge was never sufficient to overturn civilization entirely. Either the invaders would overrun the settled lands and then adopt civilized customs, or the frontier defenses would prove strong enough to hold them off. There were even long periods of peace, when the barbarian threat was negligible. It was at such times that the spontaneous ingenuity of mankind had greatest play, in politics as in everything else. But it is noteworthy that, in the end, what may be described as the ancient norm always reasserted itself, whether in Europe, the Middle East, India, or China. Military crises—civil war or barbarian invasion or both—recurred and necessitated the strengthening of government. The effort to secure a measure of peace and prosperity required the assertion of authority over vast distances, the raising of large armies, and the gathering of taxes to pay for them. These requirements in turn fostered literacy and numeracy and the emergence of what later came to be called bureaucracy, government

by official. Bureaucratic imperialism emerged again and again and spread with civilization. Barbarian challenge occasionally brought it low but never for very long. When one city or people rose to hegemony over its neighbours, it simply incorporated their bureaucracy with its own. Sumer and Babylon were conquered by Assyria; Assyria was overthrown by the Medes, in alliance with Babylon and nomad Scythians; the empire of the Medes and Persians was overthrown by Alexander of Macedon; the Macedonian successor states were conquered by Rome, which was in due course supplanted in the Middle East and North Africa by the Islāmic Caliphate of Baghdad. Conquerors came and went; life for their subjects, whether peasants or townsmen, was not much altered by anything they did, so long as the battles occurred at safely remote sites.

Bureaucratic monarchy

Nevertheless, from time to time experiments were made, for no monarchy commanded the resources to rule all its subjects directly. So long as they paid tribute punctually, local rulers and local communities were perforce left very much to govern themselves. Even if they did not pay, the effort required to mount a military operation at a distance from the imperial centre was so great that only in exceptional circumstances would it be undertaken, and even then it might not succeed, as the kings of Persia found when they launched punitive expeditions against mainland Greece at the beginning of the 5th century BC. So, in normal times, the inhabitants of the borderlands had extensive freedom of action.

Although civilization, as its advantages became clear, spread west and northwest out of Asia, bureaucratic monarchy could not easily follow it, for the sea was becoming a historical factor as important as the steppe and the great irrigable rivers. Tyre and Sidon, maritime cities of Phoenicia, had long exploited their coastal situation, not only to remain independent of the landward empires but also to push across the sea, even beyond the Strait of Gibraltar, in quest of trade. Their daughter-cities, Carthage, Utica, and Cádiz, were the first colonies, but primitive communications made it impossible for Phoenicia to rule them from afar.

GREECE

The city-state. The Phoenician example was followed by the Greeks. The Greeks were originally Indo-European nomads, who gradually made their way down to the Aegean and there took to the sea. They built on the achievements of earlier peoples, even taking over the first bureaucratic monarchy to appear on European soil: the Minoan civilization of the island of Crete, which succumbed to the invaders about 1450 BC. Continuing invasions from the north overthrew the mainland kingdoms of Mycenae, Tiryns, and Pylos in about 1200 BC. The so-called dark ages of Greece that then began lasted until the 8th century BC, by which time the Greeks had not only recovered literacy, by adapting the Phoenician alphabet, and begun to found overseas colonies, but had also brought the city-state to something near maturity. This form of government was the great political invention of classical antiquity.

Mediterranean geography is such that every little fishing village had to be able to defend itself against attack from land or sea, for outside help could not reach it easily. A man's dependence on his community, for physical as well as economic survival, was therefore obvious and complete.

The *polis*

The city (*polis*) had first claim on his labour and loyalty, a claim that was usually freely recognized. It was this reality that led Aristotle (who himself came from just such a small commonwealth, Stageira, or Stagirus) to define man as a political animal.

Coastal mountain ranges made it difficult for any community to dominate more than a few square miles of land. They also for a long time deterred the rise of an empire to federate and control all the cities. If a few of these centres nevertheless rose to imperial greatness, like Tyre and Sidon before them, it was in the main because, like their Phoenician predecessors, they traded across the sea successfully. Athens, for example, exported olive oil, silver, and pottery. The profits of this trade enabled it to build a great navy and formidable city walls. Athenian ships defeated Persia (480 BC) and won a small empire in the Aegean Sea; the combination of ships and walls enabled Athens long to defy, and nearly to defeat, its chief rival among the Greek cities, Sparta. Even after Sparta's triumph at the end of the Peloponnesian War (404 BC), Athens remained an independent sovereign state until its defeat by Philip of Macedon at the battle of Chaeronea (338 BC). In short, during the period of its prime Athens was free to make what experiments it liked in the realm of government, and to that period are owed not just the first example of successful democracy in world history but also the first investigations in political thought.

Monarchy, oligarchy, democracy. No Athenian believed that he had anything to learn from the bureaucratic monarchies of the East, which were incompatible with Greek notions of citizenship. All monarchies, indeed, seemed bound to deteriorate into such tyrannies. Self-defense necessitated that every adult male be required, and indeed be willing, to fight when called on; in return he had to be given some measure of respect and of personal autonomy—in a word, freedom. To protect that freedom, government was necessary; anarchy had no attractions for any Greek. The central question of politics, then, was the distribution of power among the citizens. Was Greek freedom best preserved and defined by the rule of the few or by that of the many? On the whole the great names favoured aristocracy, the rule of the best. Plato believed that the object of politics was virtue, and that only a few would ever thoroughly understand the science by which virtue could be attained. These trained few, then, should rule. Aristotle, his pupil, seems to have put the cultivation of the intellect highest among human goods; and he believed—quite reasonably, given the limited resources then available—that this fruit of civilization could be reaped only among a leisure class supported by the labours of the many. In return for their leisure the gentry would agree to sacrifice some of their time to the tedious business of governing, which only they would be sufficiently disinterested and well informed to do successfully. Neither of these apologies for oligarchy had any success in practice. The democrats carried the day, at any rate in Athens and its allied cities. In return for playing their parts as soldiers or sailors, ordinary Athenians insisted on controlling the government.

Views of Plato and Aristotle

The result was impressive. The people were misled by demagogues; they were intolerant enough to put Plato's master, Socrates, to death; they were envious of all personal distinction; and of their three great wars (against Persia, Sparta, and Macedon), they lost two. Furthermore, passionate devotion to the idea that Athens was the greatest of all cities, the school of Greece, and the wonder of civilization, misled them into basing their society in large part on slave labour, into wanton imperial adventure abroad, and into denying Athenian citizenship to all who were not born to it (even Aristotle), however much they contributed to the city's greatness and however much more they might have done. The foundations of Athenian democracy were narrow and shallow and fragile. But to say all this is only to say that the city could not entirely shake off the traditions of its past. Its achievement was the more remarkable for that. Seldom since has civilized humanity equalled democratic Athens, and until the last the city was satisfactorily governed by law and by popular decision. It owed its fall less to any flaw than to the overwhelming force that was mounted against it.

For to the north of Hellas proper a new power arose. Greek civilization had slowly trained and tamed the wild men of Macedon. Their king Philip forged them into a formidable army; he and his son Alexander then seized the opportunity open to them. History and geography made it impossible for the Greek cities to unite, and so they hanged separately. It seemed as if the city-state had been but a transient expedient. Henceforward Athens and Sparta would take their orders from the Great King.

ROME

The republic. As it turned out, the city-state had barely begun to display its full political potential. To the west, two non-Greek cities, Carthage and Rome, began to strug-

gle for mastery, and after the defeat of Hannibal at Zama (202 BC) Rome emerged as the strongest state in the Mediterranean.

The Greeks did not know how to classify Rome. The Greek historian Polybius, who chronicled Rome's rise, suggested that its constitution was such a success because it was a judicious blend of monarchy, aristocracy, and democracy. The Romans, a conservative, practical people, showed what they thought of such abstractions by speaking only of an unanalyzed "public thing"—*res publica*—and thus gave a new word to politics. With this focus the patriotism of the city-state reached its greatest intensity. The Romans were deeply attached to their traditions, and these traditions all taught the same lesson. For example, the legendary hero Mucius Scaevola gave his right hand to the flames to prove that there was nothing a Roman would not endure for his city. This passionate devotion to Rome's survival was tested again and again in war. All the tales of early Rome turn on battle. With dour persistence the peasants who had gathered for self-protection on the seven hills resisted every invader, fought back after every defeat, learned from all their mistakes, and even, however reluctantly and belatedly, modified their political institutions to meet the new needs of the times as they arose. Polybius was right: power in Rome was indeed shared among the people, the aristocracy, and the consuls, the executive officers of the republic who had replaced the kings. The claims of the many and of the few were fought out at election time, when the world's first clearly identifiable political parties appeared. Until Rome's late decadence, the results of elections were universally respected, and the triumphant alliance of the few and the many against the world was proclaimed in the letters blazoned on the city's buildings and battle standards: SPQR, for Senatus Populusque Romanus, the Senate and the people of Rome.

The Roman *res publica*

Like Athenian democracy, this system worked well for a surprisingly long time, and if the chief Athenian legacy was the proof that politics could be understood and debated logically and that under the right conditions democracy could work, Rome proved that the political process of competition for office and public discussion of policy was a valuable thing in itself.

Nevertheless, the Roman republic had been forged in a grim world. Its power had gradually extended over Italy in wars, always supposedly of self-defense. It is not surprising that what impressed the world most about the city was its military strength rather than its political institutions, even though the two were intimately related. As the weakness of Rome's neighbours became apparent, the Romans began to believe in their mission to rule, "to spare the conquered and war down the proud," as the poet Virgil put it. Military strength, in short, led to military adventurism. By the 1st century BC, Rome, having become a naval power as well as a military one, had conquered the whole Mediterranean basin and much of its hinterland. The strains of empire building made themselves felt. The Roman armies, no longer composed of citizens temporarily absent from the plow or the workshop but of lifetime professionals, were now loyal to their generals rather than to the state; and these generals brought on civil war as they competed to turn their foreign conquests into power at home. The population of Rome swelled, but economic growth could not keep pace, so that many citizens became paupers dependent on the public dole. The aristocrats appointed to govern the provinces saw their postings merely as opportunities to get rich quickly by pillaging their unfortunate subjects. The republic could not solve these and other problems and was in the end superseded by the monarchy of Caesar Augustus.

The empire. The bedrock of Augustus' power was his command of the legions, but he himself was a much better politician than he was a general, and he knew quite well that naked political power is as unstable as it is expensive. He reduced the military establishment as much as was prudent, laboured to turn the revolutionary faction that had supported his bid for power into a respectable new ruling class, and proclaimed the restoration of the republic. But not even Augustus could make the restoration real. With the safety of the state, questions of war and peace, and most of the business of governing the empire in the hands of a monarch, there was not enough for the Senate to do, and Augustus never went so far as to restore genuinely free elections or the organs of popular government. He kept the crowd happy with chariot races, gladiatorial contests, and the dole of bread. Nevertheless, he could not give up the attempt to legitimize his regime. Like earlier monarchs elsewhere, he called in the aid of religion, even though the religion of Rome was as republican as its constitution. Later emperors made their own divinity a tenet of the public faith.

The Augustan state

For four centuries the resemblance between Rome and the bureaucratic eastern monarchies steadily increased. Roman nationalism, Roman traditionalism, and Roman law survived as legacies that posterity would one day claim; and if nobody much believed in the constitutional shams of Augustus' day, the example of his constitutional monarchy was to prove potent at a much later period.

The age of the city-state was at last drawing to a close. The emperor Caracalla extended Roman citizenship to all subjects of the empire, but that was merely so that he could tax them more heavily. The demands of the imperial administration began to bankrupt the cities, which had previously prospered as the local organs of government under Rome. New barbarian attacks threw the empire onto the defensive, and in AD 410 Rome fell to the Visigoths.

THE MIDDLE AGES

Dissolution and instability. Seen against the background of the millennia, the fall of Rome was so commonplace an event that it is almost surprising that so much ink has been spilt in the attempt to explain it. The Visigoths were merely one among the peoples who had been dislodged from the steppe in the usual fashion. They and others, unable to crack the defenses of Sāsānian Persia or of the Roman Empire in the east (though it was a near thing), probed farther west and at length found the point of weakness they were seeking on the Alps and the Rhine.

What really needs explaining is the fact that the western empire was never restored. Elsewhere the universal throne was never vacant for long. In China, after every time of troubles, a new dynasty received "the mandate of Heaven," and a new Son of Heaven rebuilt order. For instance, in AD 304 the Huns had invaded China, and a long period of disruption followed; but at the beginning of the 7th century the T'ang dynasty took charge and began 300 years of rule. The Europeans failed to emulate this story. Justinian, the greatest of the eastern Roman emperors, reconquered large portions of the West in the 6th century, though the destruction wreaked by his soldiers made things worse rather than better. In 800, Charlemagne, king of the Franks, was actually crowned emperor of the West by the Pope. In later centuries the dynasties of Hohenstaufen and Habsburg tried to restore the empire; so, as late as the 19th century, did Napoleon Bonaparte. None of these attempts succeeded. Probably the chance was only real in the earliest period, before Europe had got used to doing without an overlord. But at that time there was never a long enough breathing space for society to regain its stability and strength. Most of the barbarian kingdoms, successor states to Rome, succumbed to later assailants. Britain fell away from the empire in the 5th century; the little kingdoms of the Angles and Saxons were just coming together as one kingdom of England when the Viking invasions began. In the 7th century the Arabs conquered North Africa, and in the 8th they took Spain and invaded Gaul. Lombards, Avars, Slavs, Bulgars, and Magyars poured into Europe from the East. Not until Otto I's victory over the Magyars at Lechfeld in 955 did these incursions cease, and not until the late 11th century was Latin Christendom more or less secure within its borders; and by then it had been without an effective emperor for more than 600 years.

Attempts at restoration

Feudalism. Various institutions had emerged to fill the gap. The church, against enormous odds, had kept the light of religion and learning alive and spread what was left of Roman civilization into Ireland, England, Central Europe, and Scandinavia. It also provided a reservoir of literacy against the day when professional government

should be possible again. The kings of the barbarians, of whom Charlemagne was the greatest, had provided military leadership and tried to acquire some of the prestige and governmental machinery of the Roman emperors. But the troublous times ensured that effective power fell to a military aristocracy. Its members called themselves *nobiles* in the Roman fashion and appropriated various late imperial titles such as *comes* (count) and *dux* (duke). But this was mere decoration. The new kings, lacking the machinery for imperial taxation, could not pay for standing armies. Besides, this was the age in which the heavily armoured cavalryman (*chevalier* in French; knight in English) dominated war. He was an independent force and thus a much less dependable instrument than a Roman legionary had been. Legally, the new masters of the soil were liegemen of the various kings (it was a maxim that every man had a lord), but in practice they could usually ignore royal claims. Europe thus fell under the rule of armoured lords; and the course of the next few hundred years gave everyone reason to wonder whether the democrats of Greece had not been right to distrust the very idea of oligarchy, for the tonic note of noble rule seemed to be almost incessant warfare.

Influence of knighthood

Even at their height the military aristocrats never had it all their own way. Strong monarchies gradually developed in England, France, and, a little later, in the Iberian Peninsula. During the heyday of the papacy (*c.* 1050–1300) the Roman Catholic Church was able to modify, if not control, baronial behaviour. Trade gradually revived and brought with it a revival of the city-state in Italy, the Rhineland, and the Low Countries, for the newly prosperous burghers could now afford to build stout walls around their towns, and it became difficult for the nobility to muster the force to besiege them successfully. Even the peasants from time to time made themselves felt in bloody uprisings. The nobility itself was far from a homogeneous or united class.

The rise of law and the nation-state. Medieval Europe, in short, was a constantly shifting kaleidoscope of political arrangements; to the extent that it ever settled down, it did so on the principle that since everybody's claim to power and property was fragile and inconsistent with everybody else's, a certain degree of mutual forbearance was necessary. This explains the great importance attached to custom or (as it was called in England) common law. Disputes were still often settled by force, especially when kings were the disputants, but the medieval European became almost as fond of litigation as he was of battle. Every great estate was hung about with quasi-permanent lawsuits, and the centralization of the church on the papal courts at Rome ensured yet more work for lawyers, the greatest of whom began to merge with the military nobility into an aristocracy of a new kind. Rights, titles, and privileges were forever being granted, revoked, and reaffirmed. Parchment deeds came to regulate men's political, social, and economic relationships at least as much as the sword did. In these ways the idea of the rule of law was reborn. By the beginning of the early modern period, legally demonstrable privileges had become the universal cement of European society. The weak were thus enabled to survive alongside the strong, and Europe, where everyone knew to which order of society he belonged, thus took on a faint resemblance to India, where the caste system was so strong as to make many of the usual functions of government quite unnecessary.

The common law

But there was a dynamism in European society that prevented it from setting permanently into this or any other pattern. This evolving Europe of privileged orders was also the Europe of rising monarchies. With many setbacks the kings clawed power to themselves; by 1500 most of them presided over bureaucracies (initially staffed by clerics) that would have impressed any Roman emperor. But monarchy could no longer claim universality. The foundations of the new monarchies were purely territorial. The kings of England, France, and Spain had enough to do to enforce their authority within the lands they had inherited or seized and to hammer their realms into some sort of uniformity. This impulse explains the wars of the English against the Welsh, Scots, and Irish; the drive of the French kings toward the Alps, the Pyrenees, and the Rhine; and the rigour of the Spanish kings in forcing Catholicism on their Jewish and Moorish subjects. Uniformity paved the way for the most characteristic governmental form of the modern world, the nation-state.

EMERGENCE OF THE MODERN WORLD

The failure of absolutism. That evolution was not easy, for kings or anybody else. The invention of gunpowder enabled kings to overbear their turbulent nobles, for cannon were exceedingly effective at demolishing the castles in which rebellious barons had formerly been quite safe. But artillery was cruelly expensive. This fact underlies the rise and persistence of the greatest political discovery of the later Middle Ages: the principle of representation. A sufficient revenue had always been one of the chief necessities of monarchy; none of the great European kingdoms had ever succeeded in securing one for long. The intractable complexities of medieval society permitted very little coercion of taxpayers; for the rest, money could be secured only by chicanery, or by selling offices or crown lands (at the price of a long-term weakening of the monarch), or by robbing the church, or by a lucky chance, such as the acquisition of the gold and silver of Mexico and Peru by the king of Spain, or by dealing, on a semi-equal footing, with parliaments (or estates, as they were more generally known).

The principle of representation

Yet the principle of parliamentary representation was of slow and uncertain growth, except perhaps in England, where the folly of the Stuart kings ensured its survival. Before then, three great occurrences, the Renaissance, the Reformation, and the discovery of America, had transformed Europe.

The impact of the Renaissance defies summary, even if its political consequences are all that need be considered. The truest symbol of its importance is the printing press. For one thing, this invention enormously increased the resources of government. Laws, for example, could be circulated far more widely and more accurately than ever before. But more important still was the fact that the printing press increased the size of the educated and literate classes. Renaissance civilization thus took a quantum jump, acquiring deeper foundations than any of its predecessors or contemporaries by calling into play the intelligence of more individuals than ever before. But the catch (from a ruler's point of view) was that this development also brought public opinion into being for the first time. Not for much longer would it be enough for kings to win the acquiescence of the nobility and the upper clergy; a new force was at work, as was acknowledged by the frantic attempts of all the monarchies to control and censor the press.

The Reformation was the eldest child of the press; it, too, had diffuse and innumerable consequences, the most important of which was the destruction of the Roman Church's claim to universality. It had always been a somewhat fraudulent claim: the pope had never actually been accepted by all the Christian bodies, even all the orthodox ones; but after Martin Luther and John Calvin, the scope of his commands was radically reduced. In the long run the consequence was the secularization of politics and administration and the introduction of some measure of religious toleration. Gradually the way became clear for rational, utilitarian considerations to shape government.

The discovery of the Americas opened a new epoch in world history. The Spanish overthrew the monarchies of the Aztecs and the Incas, thanks partly to the superior weapons that they brought with them and partly to the diseases. It was a spectacular episode, the first to proclaim that the old struggle between the steppe and the sown had been bypassed: from now on the drama of history would lie in the tension between the oceans and the land. The globe was circumnavigated for the first time; European ships, bearing explorers, traders, pirates, or men who were something of all three, penetrated every sea and harbour; and although the ancient civilizations of Islām, India, China, and Japan saw no need to alter their customs to take account of European innovations, the signal had been given for their fall.

Portuguese and Spanish explorations gave far-flung overseas empires to both countries, and perhaps as many difficulties as benefits. Other countries—France, England, the Netherlands, Sweden, Denmark—thought it both undesirable and unsafe not to seek empire themselves; the Iberian monarchies were thus involved in a perpetual struggle to defend their acquisitions. This entailed incessant expenditure, more, in the end, than the kingdoms' revenues could match. Financial weakness was one of the chief causes of the ultimate decline of Spain. But by then the inadequacies of the monarchical system had been cruelly exposed in such episodes as the revolt of the Netherlands, the defeat of the Spanish Armada by England, and, worst of all, the snail's-pace development of the colonies in the New World. Charles V and Philip II were as able as all but a few monarchs in history; but they could not overcome the structural weaknesses of hereditary monarchy. There was no mechanism by which they could devolve their most crushing duties on their ministers, so government moved slowly, if at all. As lawful sovereigns they were bound by the customs of their numerous realms, which frequently blocked necessary measures. They were unable to guarantee that their heirs would be their equals. The same difficulties eventually ruined the French monarchy. The only remedy discoverable within the system was for the king in effect to abdicate in favour of a chief minister. Unfortunately a man equal to the task was seldom found, and anyway no minister, however great, was ever safe from the constant intrigues and conspiracies of disgruntled courtiers. Problems tended to accumulate until they became unmanageable.

Weaknesses of monarchy

Representation and constitutional monarchy. Meanwhile, in England, the rise of Parliament introduced a republican, if not a democratic, element into the workings of one of Europe's oldest kingdoms.

The republican tradition had never quite died out. The Dutch had emerged from their long struggle against Catholic Spain clinging triumphantly to their new religion and their ancient constitution, a somewhat ramshackle federation known as the United Provinces. Switzerland was another medieval confederation; Venice and Genoa were rigidly oligarchical republics. What set England apart was that everywhere else in monarchical Europe the medieval representative institutions—the États-Généraux in France, the Cortes in Spain, the various diets of the Germanic countries—fell into oblivion. In England alone the expense of naval war, the incompetence of the kings, the exhaustion of all other sources of finance, and the absence of a standing army brought about the continuance and strengthening of Parliament. The climax came under William III, a Dutchman who was quite content to let Parliament take an unprecedentedly large share in government so long as it voted money for his war against Louis XIV of France. He conceded, in short, full power of the purse to the House of Commons, and before long it became a maxim of the dominant Whig Party that no man could be legally taxed without his own consent or that of his representatives. A radically new age had dawned.

The Whig system was called constitutional monarchy. Disputes about the right and legal way of governing England had raged throughout the 17th century. Civil wars and revolutions had been accompanied, perhaps in part caused, by endless rummaging through old manuscripts and a nearly ceaseless war of printed pamphlets. The increasingly rationalist temper of the times, exemplified in the works of John Locke, finally buried some of the more blatantly mythological theories of government, such as that of the divine right of kings; and after the flight of King James II in 1688, Parliament finally settled the issues that had so vexed the country by passing a series of measures that gave England a written fundamental law for the first time. Henceforth the country was to be ruled by a partnership between King and Parliament (in practice, between the king and an oligarchy of country gentlemen who controlled most parliamentary elections); and if many Englishmen looked with distaste on the squabbles of party politics, which were the sordid result of this arrangement, few could propose a plausible alternative. Tories drank toasts to the King over the water; republicans published more pamphlets; and Sir Robert Walpole ruled for 21 years as the first British prime minister.

Rise of the British Parliament

The secret of Walpole's strength lay in his ability simultaneously to please the King, to give the country sound government finances, and to command a majority in the houses of Parliament. He performed this last trick partly by giving out sinecures, salaries, and titles to his supporters, partly by his superiority in debate, and partly by exploiting Whig fears of Tories and Roman Catholics. These three elements—party interest, practical decision making, and party ideology—have in one form or another come to dominate most modern political systems where brute force is at a discount.

Even after Walpole's fall his arrangements continued. They were vindicated by the Seven Years' War (1756–63), when Britain defeated both the French and Spanish empires and emerged predominant in every ocean and (especially) in North America. Immediately afterward republican ideology found its classical expression.

The American and French republics. The limited British monarchy found it little easier to govern a sea-borne empire than did the kings of France and Spain. If Britain's American colonies were to grow in population and riches, so as to become sources of strength to the empire, not military and financial liabilities, they had to be given a substantial measure of religious, economic, and political independence, and the gift could not be revoked. Once British policy had created a chain of more or less self-governing communities along the Atlantic seaboard of North America—communities much like the city-states of old—it could not undo its own work, even when it found its clients unreasonable, small-minded, and recalcitrant. Thus, when the British government attempted to impose tighter rule from London, the old empire broke down in bickering about taxation and, in riot, rebellion, and civil war—in short, in the American Revolution. From 1775 to 1783 the Anglo-Americans fought with determination and good luck against their former overlord, King George III; and in 1776 their leaders determined to be free of him and the British Parliament forever. The principles on which they meant to found a new commonwealth were expounded in the Declaration of Independence:

The Declaration of Independence

> . . . We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain inalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. —That, to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed. . . .

The application of these principles was more difficult than their enunciation, and it took the Americans 12 years to get themselves a suitable form of government. When they did adopt a new constitution it served them so well that it is still operative. This durability is not unconnected with the fact that the Constitution opened the door to modern liberal democracy. "The consent of the governed" was agreed to be the key to governmental legitimacy, and in practice the phrase rapidly came to mean "the consent of the majority." The principle of representation was embodied in the Constitution (the first article of which was entirely devoted to the establishment of Congress, the American parliament); this implied that there was no limit to the potential size of a successful republic. From Plato to Rousseau, theorists had agreed that democracies had to be small, because by definition all their citizens had to be able to give their consent in person. This restriction was alone almost sufficient to explain the recurrent failures of republican government down the ages. Now it had been evaded.

Yet its example might have had little effect on Europe but for the French Revolution of 1789. The French had helped the Americans to beat the British, but the effort had been too much in the end for the monarchy's finances. To avert state bankruptcy the États-Généraux were summoned for the first time in 175 years, and soon the whole government had been turned upside down. The French repudiated the divine right of kings, the ascendancy of the nobility, the privileges of the Roman Catholic Church, and the regional structure of old France, and, in the end, they set up a republic and cut off the former king's head.

Unfortunately for peace, in destroying the monarchy the Revolution also crowned its centuries-old labours. The kings had created the French state; the Revolution made it stronger than ever. The kings had united their subjects in the quest for glory; now the nation made the quest its own. In the name of rationality, liberty, and equality (fraternity was not a foremost concern), France again went to war. The Revolution had brought the new invention, the nation-state, to maturity, and soon it proved capable of conquering the Continent, for everywhere French armies went the revolutionary creed went, too.

The nation-state

In all this the French Revolution was giving expression to a general longing for government devoted to the greatest happiness of the greatest number. But there was also considerable resistance, which increased as time went on, to receiving the benefits of modern government at the hands of the French. Besides, the facts of demography began to tilt against the Revolution, as population growth in Britain and Germany accelerated and that of France slowed down. The century that began with the victories of Marengo, Austerlitz, and Jena ended with the Third Republic nervously on the defensive and French society still bitterly at odds with itself.

Yet, on the whole, the work of the French Revolution survived. However many changes of regime France endured (seven between 1814 and 1870), its institutions had been thoroughly democratized, and the underlying drift of events steadily reinforced this achievement. By mid-century universal manhood suffrage had been introduced, putting France in this respect on the same footing as the United States. Britain, pursuing its own historical logic, evolved in much the same way; its oligarchs slowly and ungraciously consented to share political power with other classes, rather than lose it altogether. By the end of the century manhood suffrage was clearly at hand in Britain, too, and women would not much longer be denied the vote. Smaller European countries took the same course. And everywhere (in America as well as western Europe) the representative principle combined with the necessity of government to produce the modern political party. Elections could only be won by organized factions; politicians could only attain or retain power by winning elections. Permanent parties resulted. The Industrial Revolution and continuing population growth made an elaborate state apparatus more and more necessary. The spread of education and prosperity made more and more citizens feel fully equal to taking part in politics, whether as voters or statesmen. Modern government in the West thus defined itself as a blend of bureaucracy, party politics, and passionate individualism, the whole held together by the cement of an equally passionate nationalism.

Political parties

Nationalism and imperialism. The kingdom of Prussia and the empires of Austria and Russia readily learned from the French Revolution that it was necessary to rationalize government. They had been struggling along that path even before 1789. Carrying out the necessary changes proved exceedingly difficult (Russia, which in some ways was more like ancient Egypt than a modern country, made far too few changes until far too late), but gradually they were managed. Meanwhile the libertarian and egalitarian components of the revolutionary legacy were rigidly resisted. The great dynasts, and the military aristocracies that supported them, had no intention of admitting their obsolescence. Though they were forced to make limited concessions to the liberal spirit between 1789 and World War I, the autocratic citadel of their power was never surrendered. Instead, the myth of the nation was adopted to reinforce the authority of the state.

Nationalism intensified the competitiveness that had always been a part of the European state system. Peoples, it emerged, could be as touchy about their prestige as monarchs. But for a hundred years an uneasy peace prevailed in Europe, leaving the powers free to pursue interests in other parts of the world. Asia and Africa thus came to feel the full impact of European expansion, as the Americas had felt it before. Only the Japanese proved to have the resources to adapt successfully to the new ways, taking what suited them and rejecting the rest. They kept their millennial sacred monarchy but modernized the armed forces. In 1895 they fought and won a war against China, which was sliding into chaos, and in 1905 they defeated a great power, Russia. But Japan was wholly exceptional. Elsewhere, European power was irresistible. Britain gave up the attempt to govern overseas settlements of its own people directly—the experiment had proved fatal in America and nearly so in Canada—but retained indirect authority over them, in part through the useful device of its own sacred monarchy. And it had no scruple about assuming direct rule over more and more non-British peoples. France, Germany, and the United States eagerly followed the example; The Netherlands, Spain, and Portugal clung to what they had, though the last two suffered great imperial losses as Mexico, Brazil, and others shook off colonial rule. It seemed that before long the whole world would be ruled by half a dozen powers.

It did not turn out so, or not for long. The problem of governmental legitimacy in central, eastern, and southern Europe was too explosive. The obstinate conservatism of the dynasts proved fatal to more than monarchy. There were too many who regarded the empires as unacceptable, either because they were the instruments of class oppression, or because they embodied foreign rule, or both. And the romantic tradition of the French Revolution—the fall of the Bastille, the Reign of Terror, the Jacobin dictatorship—helped to drive many of these critics into violent rebellion, permanent conspiracy, and corrosive cynicism about the claims of authority. Authority itself, corrupted by power and at the same time gnawingly aware of its own fragility, embarked on more and more risky militarist adventures. The upshot was the war of 1914, the revolutions that followed it, and the new model government that emerged in Russia at the end of the period.

Legitimacy and upheaval

20TH-CENTURY MODELS

The Soviet state. In theory the Soviet Union was and is a democracy. In practice it has proved to be one of the most effective oligarchies in history. Lenin and his followers won power in the turmoil of revolutionary Russia because they were abler and more unscrupulous than anyone else. They retained and increased their power by force, but they argued that the social theories of Karl Marx, as developed by Lenin, are of universal, permanent, and all-sufficient validity; that the leadership of the Communist Party has a unique understanding of these theories; and that its will can never legitimately be resisted. All institutions of the Soviet state are devised primarily to assure the untrammelled power of the party, and no methods have been spurned, from mass starvation to the suppression of works of art, in furthering this aim. Even the extraordinary economic and military achievements of the regime have been secondary to its overall purpose.

From the first the Soviet model found many imitators. Lenin's strictly disciplined revolutionary party, the only morality of which was unswerving obedience to the leader, was a particularly attractive example to those intent on seizing power in a continent much weakened by the war. Benito Mussolini of Italy modelled his Fascist Party on the Leninists; Adolf Hitler of Germany copied Mussolini. Both men won power, which they criminally abused, in conditions of social and political chaos like those Lenin had exploited. In the Soviet Union itself, Lenin's successor and disciple, Joseph Stalin, outdid his master in building up his power by mass terror and party discipline. Between them these three dictators brought about World War II, after which Stalin was able to extend the Soviet model directly to the eastern European countries that victory enabled him to dominate. Since his death in 1953 his empire has lurched from crisis to crisis, never resolving its basic dilemma. The terror, the lies, the regimentation are less insistent than they were in the 1930s and '40s, but the regime dares not dispense with them altogether; yet while they persist there can be no rational and effective program for tackling the Soviet Union's many problems, of which all-pervasive corruption is probably the worst.

Stalinist totalitarianism

Liberal democracy. Government in the West has evolved along more reassuring lines. The democratic dogma is everywhere in the ascendant and has brought with it greater influence for the working classes, for women, for

non-European races, and for small states. Liberalism has a unique claim to its credit: no Western democracy has ever fought a war against another. The muddled processes of open debate and decision have produced an economic order that has proved vastly more productive than the command economies of the East and has brought greatly improved living standards for almost all. Free elections have meant that bad, or at any rate unpopular, governments have been regularly dismissed by the voters with historically unparalleled ease and peacefulness. But Western democracy, however perfect its forms (and nowhere are they entirely consistent with its principles), always has problems on its hands that may prove too much for it. It could not avert the outbreak of two world wars, and a third has been averted so far more through terror of nuclear weapons than by democratic wisdom. Class conflicts are muted rather than resolved. Nationalism still distorts voters' judgments in matters of foreign policy; greed misleads them over economic policy. Demagogues are as much a menace as they ever were in ancient Athens, and many politicians are personally corrupt.

The great experiment of European imperialism has collapsed. The two world wars robbed the powers of the will and the means to maintain overseas rule. Unfortunately the empires have not been succeeded by new governmental forms fully capable of dealing with the problems of technologically backward, overpopulated, culturally premodern societies. In the Muslim world the idea of the Islāmic republic has arisen. In Iran it amounts to an attempt to wed church and state, religion and government, as indissolubly as they were in the age of the first caliphs, only now the religion is to be Shi'ah Islām, and the state a modern republic. If the experiment proves a success it may well have as widespread an impact within the Muslim world as Leninist Communism has had on revolutionaries everywhere; but by the nature of the case it cannot spread outside regions of Islāmic dominance. In India a regime of more or less democratic nationalists endeavours to overcome a pervasive regionalism and social stratification that have been part of the subcontinent for more than 2,000 years. China, which during the lifetime of Mao Tse-tung wholeheartedly followed the Leninist model, began after his death to draw back. Japan has adapted to Western notions of capitalism and parliamentarism without contributing any fresh ideas. Elsewhere in the world the commonest expedient adopted has been dictatorship, usually military, in which the ancient tradition of autocracy reasserts itself; but it is autocracy without its ancient stability, and of all current political forms it seems by far the least likely to deal effectively with the universal enemies: hunger, war, poverty, disease, waste, violence. Some thinkers believe that only a form of world government can make decisive headway against these evils, but no one has yet suggested either how world government can be set up without a world war or how, if such a government did somehow come peacefully into existence, it could be organized so as to be worthy of its name. Even effective cooperation among national governments often seems unattainable, as the story of such international bodies as already exist demonstrates all too sadly.

The Islāmic republic

For the time being the world exists in uneasy equipoise. The United States and the Soviet Union, dedicated rivals, are too big and willful to subordinate themselves to international organizations. The nation-state has, with few exceptions, become the political norm. The issues that once made the distinctions between monarchies and republics so important have withered into insignificance. Everywhere the growing complexity of life has necessitated, and technology has made possible, an immense expansion of the state, so that it seems at times almost as if the Italian economist and sociologist Vilfredo Pareto, who believed that all government is but a facade for oligarchy, was right. But even if so, problems still remain for which mere oligarchy is no answer. The incompatible claims of the city-states ruined ancient Greece; the modern world may yet be ruined by the incompatible claims of the nations. If man, the political animal, is to save himself and his civilizations, he cannot yet rest from seeking new forms of government to meet the ever-new needs of his times.

BIBLIOGRAPHY. Classical texts on governmental forms are widely available in numerous editions. They include PLATO, *The Republic;* ARISTOTLE, *Politics;* NICCOLÒ MACHIAVELLI, *The Prince* and *The Discourses;* THOMAS HOBBES, *Leviathan;* JOHN LOCKE, *Second Treatise of Government;* MONTESQUIEU, *The Spirit of Laws;* JEAN-JACQUES ROUSSEAU, *The Social Contract;* THOMAS JEFFERSON, *Declaration of Independence;* ALEXANDER HAMILTON, JAMES MADISON, and JOHN JAY, *The Federalist;* EDMUND BURKE, *Reflections on the Revolution in France;* THOMAS PAINE, *Rights of Man;* ALEXIS DE TOCQUEVILLE, *Democracy in America;* WALTER BAGEHOT, *The English Constitution;* JAMES BRYCE, *The American Commonwealth;* FRIEDRICH ENGELS, *The Origin of the Family, Private Property, and the State;* V.I. LENIN, *The State and Revolution.*

For a subject of this nature, a solid background in world history is absolutely necessary, and WILLIAM H. MCNEILL, *A World History,* 3rd ed. (1979), is an excellent introduction. GORDON CHILDE, *What Happened in History,* rev. ed. (1954, reprinted 1982), is a classic survey of the contribution of archaeology to our understanding of prehistory and the ancient world. For the development of governmental forms in Greece see *The Cambridge Ancient History,* 3rd ed. (1970–); and N.G.L. HAMMOND, *A History of Greece to 332 B.C.,* 2nd ed. (1967, reprinted 1981). A.H.M. JONES, *Athenian Democracy* (1957, reprinted 1975), is indispensable, particularly as a corrective to Plato, Aristotle, and Thucydides. The most exciting work about the Romans written since Gibbon is RONALD SYME, *The Roman Revolution* (1939, reprinted 1974); those wanting more general accounts of the Romans may turn to J.P.V.D. BALSDON (ed.), *The Romans* (1966); and H.H. SCULLARD, *From the Gracchi to Nero,* 5th ed. (1982). A.H.M. JONES, *The Later Roman Empire,* 2 vol. (1964), is the most authoritative account of the fall of the empire.

The best introduction to the medieval papacy is WALTER ULLMANN, *A Short History of the Papacy in the Middle Ages* (1972). The same author's *Growth of Papal Government in the Middle Ages,* 3rd ed. (1970), links ideas and institutions but makes few concessions to beginners. MARC BLOCH, *Feudal Society* (1961, reissued 1974; originally published in French, 1939), is an indispensable study of its subject.

The development of political thought from the Renaissance to the 19th century is well presented in JOHN P. PLAMENATZ, *Man and Society,* 2 vol. (1963, reprinted 1972–74). An attempt to trace modern government back to its origins is demonstrated in BARRINGTON MOORE, JR., *Social Origins of Dictatorship and Democracy* (1966). See also PERRY ANDERSON, *Lineages of the Absolutist State* (1974); and FRANCO VENTURI, *Utopia and Reform in the Enlightenment* (1970; originally published in Italian, 1970).

Works on governmental forms of a general, comparative nature include A. GOODWIN (ed.), *The European Nobility in the Eighteenth Century,* 2nd ed. (1967); A. LAWRENCE LOWELL, *Governments and Parties in Continental Europe,* 2 vol. (1896, reissued 1970); EUGENE N. ANDERSON and PAULINE R. ANDERSON, *Political Institutions and Social Change in Continental Europe in the Nineteenth Century* (1967); MICHAEL OAKESHOTT (ed.), *The Social and Political Doctrines of Contemporary Europe* (1939, reissued 1949); and S.E. FINER, *Comparative Government* (1970, reissued 1974).

Useful studies of particular states include J.H. PLUMB, *The Growth of Political Stability in England, 1675–1725* (1967, reissued 1980; U.S. title, *The Origins of Political Stability, England, 1675–1725);* and ÉLIE HALÉVI, *A History of the English People in the Nineteenth Century,* 6 vol. (1924–34; originally published in French, 5 vol., 1913–32); for France, C.B.A. BEHRENS, *The Ancien Régime* (1967, reissued 1976); GEORGE LEFEBVRE, *The French Revolution,* 2 vol. (1962–64; originally published in French, 1930); and D.W. BROGAN, *The French Nation from Napoleon to Pétain* (1957, reissued 1961); for Germany, HANS ROSENBERG, *Bureaucracy, Aristocracy, and Autocracy: The Prussian Experience, 1660–1815* (1958, reprinted 1968); A.J. NICHOLLS, *Weimar and the Rise of Hitler,* 2nd ed. (1979); FRANZ NEUMANN, *Behemoth,* 2nd ed. (1944, reissued 1963); and MARTIN BROSZAT, *The Hitler State: The Foundation and Development of the Internal Structure of the Third Reich* (1981; originally published in German, 1969); for Italy, DENIS MACK SMITH, *Italy: A Modern History,* new rev. ed. (1969); and S.J. WOOLF (ed.), *The Nature of Fascism* (1968); for the Soviet Union, JEROME BLUM, *Lord and Peasant in Russia: From the Ninth to the Nineteenth Century* (1961, reprinted 1971); LEONARD SCHAPIRO, *The Origin of the Communist Autocracy: Political Opposition in the Soviet State, First Phase 1917–1922,* 2nd ed. (1977); and DEREK J.R. SCOTT, *Russian Political Institutions,* 4th ed. (1969); and for the United States, RICHARD HOFSTADTER, *The American Political Tradition and the Men Who Made It* (1948, reprinted 1974), and *The Idea of a Party System: The Rise of Legitimate Opposition in the United States, 1780–1840* (1969).

(H.Br.)

Government Finance

Government finance refers to the wide range of activities undertaken by governments in financial and economic matters. In earlier times, a government's financial activities were limited to raising funds, whether from taxes or other levies, to support its wars and its other activities. For centuries sovereigns were allowed to spend such funds as they chose; the control of expenditures came much later, not becoming a matter of concern for legislatures until the late 17th and 18th centuries. The tradition grew up that expenditures must flow from appropriations, and appropriations were to be made for specific purposes. Out of this evolved the more formal governmental budgets of today.

In the 20th century, however, the significance of a budget extends far beyond these accounting functions and the monitoring and control of expenditure. As the proportion of national income that is controlled by the various levels of government has increased, so has the importance of the budget for the general working of the economy. If a government raises more money than it spends (creating a surplus), the overall level of spending in the economy—along with personal income and employment—may be reduced. Conversely, if the government spends more than it raises, which creates a deficit, fiscal and monetary expansion is implied, which may influence inflation and growth. Thus budgets have come to serve as instruments of economic management, used to stabilize or stimulate the economy. The composition of the budget also has other economic effects. The most important of these effects are brought about by its allocative function (deciding what is supplied by the public sector and to whom) and its distributive function (deciding who gets what and how income is distributed and redistributed). The national budget, then, has become the most important instrument of a nation's economic management, and in many countries the presentation of the new budget has become synonymous with the primary annual assessment and statement of economic policy. For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 531–536.

This article is divided into the following sections:

The governmental budget 148
- Role of the budget 148
  - Traditional functions
  - Modern functions
- The accounting functions of the budget 149
- Alternative approaches to the budget 149
  - Administrative budget
  - Current and capital budget
  - Cash and unified budgets
  - Program budgeting and zero-base budgeting
  - Full-employment budget
  - Value for money measurements
  - Budgetary planning: cash, volume, and cost terms

Components of the budget 152
- Expenditure 152
  - Composition of public expenditure
  - Growth of public expenditure
  - Problems of public expenditure control
- Revenue 153
  - Taxation
  - Sale of goods and services
- Government borrowing 154
  - Forms of public debt
  - Economics of government borrowing
  - Limitations on public sector debt
  - Wartime finance
  - Evolution of government borrowing
  - Sovereign debt

The budgetary process 159
- Government versus private sector budgeting 159
- Selected national budgetary procedures 159
  - The United States
  - The United Kingdom
  - West Germany
  - Japan
  - The Soviet Union
- Budgets of other levels of government 161
  - State and local budgets in the United States
  - Local government finance in the United Kingdom
  - Budget of the European Economic Community (EEC)

The economic functions of government 162
- The allocative function 162
  - Public goods
  - Merit goods
  - Cost-benefit analysis
  - Public ownership and privatization
  - Other forms of government intervention
- The stabilization function 163
  - History of stabilization policy
  - Stabilization theory
  - Fiscal policy
  - Monetary policy
  - Stabilization policy problems
  - Experience in selected countries
- The distributive function 167
  - Incidence of taxation and expenditure
  - Arguments for income redistribution

Bibliography 168

The governmental budget

A governmental budget is the forecast by a government of its expenditures and revenues for a specific period of time. In national finance, the period covered by a budget is usually a year, known as a financial or fiscal year, which may or may not correspond with the calendar year. The word budget is derived from the Old French *bougette* ("little bag"). When the British chancellor of the Exchequer makes his annual financial statement, he is said to "open" his budget, or receptacle of documents and accounts.

ROLE OF THE BUDGET

Traditional functions. Government budgetary institutions in the West grew up largely as a result of the struggle for power between the legislative and executive branches of government. With the decline of the feudal system, it became necessary for kings and princes to obtain resources for their ventures from taxation rather than dues.

Early budgeting in the West

With the disappearance of the old feudal bonds, taxpayers demanded to be consulted before they were taxed. In England this was written into Magna Carta (1216), which stated:

> No scutage or aid shall be imposed in our kingdom unless by common counsel of our kingdom, except for ransoming our person, for making our eldest son a knight, and for once marrying our eldest daughter, and for these only a reasonable aid shall be levied.

This related to taxes only, not expenditures. For centuries Parliament seemed content to restrict the amounts that the sovereign levied while letting him spend the money as he pleased. Only after the controversies of the 17th century culminated in the Revolution of 1688 and the Bill of Rights did Parliament extend its concern from taxation to the question of expenditure control.

The histories of many countries have turned on financial crises. In France, for instance, the struggle between the monarchy and the nobility over control of tax revenues

was one of the causes of the Revolution of 1789 that led to the overthrow of both the monarchy and the nobility.

The U.S. budget system also evolved out of controversy. In the early days of the republic there was a dispute between Alexander Hamilton and Thomas Jefferson as to the amount of discretion that the executive branch should exercise in the spending of public funds. Jefferson's victory enabled Congress to assert its authority by making appropriations so highly specific as to hinder executive action. Had Hamilton won, the treasury would have attained extraordinary power in relation both to Congress and to the president.

Modern functions. In the 20th century a high proportion of economic activity is controlled, directly or indirectly, by various levels of government (federal, or central, state, local, etc.). Thus the budget has taken on a number of other functions as well as the simple monitoring of the overall revenue and expenditure of government. Expenditure programs are now planned in considerable detail, but the sheer scale of public spending raises major control problems, and varying systems of control have been tried in different countries. Taxation is used not only to raise revenue but also to redistribute income and to encourage or discourage certain activities. Government borrowing, in order to finance recurring deficits or wars, is so substantial that budgetary policy has important effects on capital markets and on interest and credit generally. Because the budget is now so important to national economies, a number of different procedures for deciding on the structure of the budget have been developed, and these vary considerably between countries. In some, the United Kingdom, for example, most planning is carried out in secret by ministers and civil servants, and public and parliamentary debate is minimal; while in others, the United States, for example, there is lengthy debate during which the budget can be changed significantly. The different levels of government complicate the budgetary process with differing spheres of influence and control over particular items of expenditure.

Varieties of budget structure

The budget has also come to be used to achieve specific goals of economic policy. It was long recognized that government borrowing could have important effects on the rest of the economy. As the scale of government activity increased, the levels of expenditure and taxation were seen to have substantial direct effects on the total demand for goods and services in the economy. This raised the possibility that by changing these levels the government could use its fiscal policy to achieve full employment and reduce economic fluctuations. This stabilization function has been used by many countries, with varying degrees of success, to expand the economy out of recession and to control inflationary pressures. In the United Kingdom, for example, postwar policy involved a sequence of "stop-go" moves by government for stabilization; unfortunately these often occurred too late and had unintended destabilizing effects.

The stabilization function

As well as affecting the overall economy, the budget may have significant (intended and unintended) effects in specific areas. Taxes affect incentives to work or to consume, while taxes, benefits, and expenditures all affect the distribution of income. In this manner, budgets, particularly those that cause major changes, have considerable political as well as economic impact.

THE ACCOUNTING FUNCTIONS OF THE BUDGET

Traditionally the budget is presented to allow scrutiny (by taxpayers, voters, and the legislature) of the resources raised by government and the uses to which these will be put. The publication of a budget thus performs the role of generating accountability for the actions of government at various levels. Historically, the focus of budgets has been to ensure that expenditures and revenues are properly authorized; more recently, the budget has been developed as a framework within which complex decisions on the allocation of resources can be made more effectively.

ALTERNATIVE APPROACHES TO THE BUDGET

In order to deal with the increasing complexity of government's role, most countries have experimented with a variety of forms for the budget and its presentation. Among the more important of these are the administrative budget, the current and capital budget, program and zero-base budgeting, and the full-employment budget. The variety of budgeting methods is extended to the types of efficiency measures used to increase value for money and to the alternative methods of projecting expenditures in cash, volume, and cost terms.

Administrative budget. The traditional administrative budget contains the executive's recommendations concerning the raising of what Magna Carta referred to as "scutage or aid" and the disposal of it for purposes of government. This kind of budget is designed to control expenditure; accordingly, it emphasizes the salaries and tasks of civil servants rather than the results that they are supposed to achieve. The control objective of the administrative budget naturally gives rise to the doctrine that the budget should be balanced. Deficits imply irresponsibility. Surpluses imply the imposition of unwarranted tax burdens on the public.

The limitation of the administrative budget is that some important items receive less than adequate attention or are excluded from it entirely. Military procurement is one example. Neither budget offices nor appropriations committees are well equipped to scrutinize the actual procurement of ships or aircraft. Consequently, in most countries large expenditures on military items are often treated perfunctorily while the activities of civil servants receive inordinate amounts of attention. The basic weakness of the administrative budget is that it is principally concerned with whether expenditure has been properly authorized, rather than whether money has been well spent.

Weakness of the administrative budget

Moreover, the administrative budget often excludes trust funds used to finance contributory old-age and unemployment insurance; taxes are paid directly into the funds and disbursements made out of them. The theory is that the government acts as trustee for the public and that the public is protected by having its social security taxes put in a separate fund. Many countries have adopted this idea of "social insurance"; it formed the heart of Bismarck's social policy for Germany in the 1870s and of the British welfare state, founded in 1948. In most cases, however, the attempt to generate a distinct fund has failed, and "contributions" have become just another tax with expenditures on, for example, retirement pensions paid irrespective of the resources available to the fund.

Other items may be included in the budget on a net rather than a gross basis. For instance, the total receipts and expenditures of the post office or other commercial activities of the public sector usually do not appear; only the deficit or surplus does. This is justified by the theory that, first, business management is not well performed by legislative committees and, second, that so long as a business undertaking pays its way, its conduct is not a matter of public concern. The problem is that the distinction between commercial and noncommercial activities is often arbitrarily made.

Current and capital budget. The administrative budget traditionally deals only with current expenditures; in many countries, some items are regarded as inappropriate for inclusion because they finance capital expenditures or are loans to other public bodies. Such items are then included "below the line," and the traditional concept of budget balance is not applied to them; instead, it is regarded as permissible to finance them by borrowing.

"Below the line" items

Direct public works or investment in nationalized industries are regarded by most countries as suitable for loan financing on the ground that they are productive assets that will yield a revenue sufficient to cover their cost. They may do so either directly, as in the case of a toll highway, or indirectly by increasing the general economic welfare, as in the case of a free highway. If, however, there is no market in which the output of a public activity is sold, there can be no objective test of its value. Hence, governments are often tempted to classify expenditure on such assets as capital items that yield a social but no economic return (*e.g.*, free playgrounds) or a lower economic return than any private sector institution would accept (as in government support for declining industries). For this

reason, distinctions between current and capital expenditures in public accounts are often viewed with suspicion.

This suspicion may be increased where, as is often the case, the rules for what is regarded as current or capital are rather indistinct. Moreover, governments have been reluctant to adopt the systematic distinction between current and capital items, or between cash flows and profit and loss accounts, or to construct a balance sheet, even though these mechanisms of monitoring receipts and expenditures are universal in private sector accounting. The federal government of the United States, for example, has resisted the idea of a capital budget, even though there was strong pressure for one in the 1930s when economists and politicians wanted to legitimize the government deficit. Among U.S. state and municipal governments, however, loan financing of public works is the regular practice for two reasons. First, those bodies are usually unable to finance their projects by current taxation; second, they do not want to finance them because the projects are generally of a long-term nature.

U.S. resistance to a capital budget

Most national governments have become accustomed to thinking in terms of national economic policies in which the amount of borrowing to be undertaken depends on current requirements for stability and growth. This makes capital budgeting less attractive, particularly if the government wishes to use the budget to supplement the national flow of savings. The more need there is to increase saving, the smaller should be the amount of government borrowing. On the other hand, government borrowing is justified when private savings tend to exceed private capital requirements.

This lack of explicit monitoring for the capital position of governments can have serious consequences when the government unwittingly takes on large liabilities or uses capital assets to finance current expenditures. Examples are provided by the growing problems in some countries in financing generous state pension schemes and the wasting of assets such as oil reserves.

Cash and unified budgets. Faced with the increasing complexity of government activities, many countries have fallen back on the idea of the cash budget. This has the merits of simplicity and comprehensiveness. As used in the United States, it presents total payments by the federal government to and from the public (including other levels of government). It is thus similar to the cash flow account of a modern business. Trust fund expenditures and receipts are included, as well as cash payments and receipts involved in loan transactions. Government business undertakings such as the post office, however, are still included on a net basis.

In the United Kingdom all public expenditure planning is now performed on a cash basis, and many programs are "cash limited," whatever the level of inflation. This procedure, to which the United Kingdom moved in 1976, is justified on the grounds that such treatment helps to control inflationary pressures and exerts stricter control than, for example, planning in volume terms.

The cash budget suffers from the defect that it is not directly tied to government decision making. Liabilities incurred do not synchronize completely with payments. This is because government expenditures result from appropriations and other forms of commitments; cash expenditures may follow appropriations and other commitments of money only after a considerable lag, notably in the case of construction and procurement. Appropriations relate to actions in the future. Expenditures result from past decisions. Both kinds of information are needed for a complete appraisal.

Time lags in cash budgets

The U.S. government, in an effort to reduce public confusion over the large variety of budgetary concepts, has adopted a so-called unified budget concept that is more logical than the cash budget but differs from it only in some details that do not materially affect the budget aggregates. The unified budget differs from the traditional administrative budget in two main ways: it includes the receipts and outlays of most funds, and it eliminates interagency transfers.

Program budgeting and zero-base budgeting. Traditionally, government expenditures have been considered as inputs rather than outputs. This is because, in the classical 19th-century conception, the well-run government does not produce a marketable output. The program budget derives from this concept; it attempts, however, to classify expenditures in terms of the outputs to which they are devoted. For example, a traditional school budget would categorize expenditures in terms of teachers, books, and buildings; what came out of the process would be left to the reader's intuition or experience. The program budget, in contrast, attempts to assign expenditures to specific outputs, categorizing them according to numbers of children completing various programs.

In government, budgets have traditionally been constructed according to departments and agencies of government. This may be justified on historical or administrative grounds, but it does not necessarily correspond to the structure of activity. Every country organizes the civilian and military components of its foreign policy in separate departments, but this is frequently a serious obstacle to effective policy-making. Again, the requirements of good administration suggest that there should be a single department of agriculture. But that department's activities impinge on those of others, in both domestic and foreign policy. A budget constructed according to actual programs would cut across departmental boundaries.

Program budgeting is an attempt to apply the economics of choice to public decision making. Its basic assumption is that explicit choice among alternative courses of action leads to better results than do other methods of decision making. At the highest governmental levels difficult choices must be made that involve the use of a portion of the nation's resources. But the same principles apply to decision making at lower levels. The problem of allocating resources within a specific field, such as health or education, is conceptually similar to that faced in drawing up the national budget.

Basic assumption of program budgeting

Program budgeting also takes account of the time dimension in many government programs. New undertakings often take time to come into operation. A typical new program may have to pass through a research and development phase and an investment or construction phase before it reaches the operating phase. Alternative programs may differ considerably in this respect. The process of choosing among alternatives frequently involves trading the present against the future. One alternative may require 10 years before it yields results; another may yield smaller results but more quickly. The kinds of choices made in government often involve alternatives that cannot be measured in terms of market value. For this reason governmental decisions involve much more uncertainty than do most business decisions.

A governmental program must therefore be frequently revised in the light of unfolding circumstances. Indeed, every year should be thought of as the first year of a new program. Pervasive uncertainty also requires a high degree of flexibility and a capacity for program revision. A number of options should be held open, particularly in the development phase. Even though this may appear costly, it is less costly than commitment to a design that proves to be inappropriate because of circumstances that could not be foreseen in the early stages.

In most countries the usual procedure for deciding on government expenditure in a forthcoming year has been to assume that existing expenditure was appropriate and then to decide on incremental expenditure for each program. Such an approach means, however, that the change is likely to increase, rather than decrease, expenditure and that little attention is paid to what the full existing program actually accomplishes.

In the late 1970s many countries recognized that the steady growth in public expenditure was putting a strain on their economies, and they attempted to curtail the growth. The Jimmy Carter administration in the United States, although planning for a steep rise in expenditure as a proportion of gross domestic product (GDP), also attempted to introduce the concept of "zero-base budgeting," whereby the entire government program, not just its incremental parts, was to be evaluated each year. This idea, which involved considerable changes to existing pro-

Concept of "zero-base budgeting"

cedures, was applied to some programs on a selective basis but never had the impact its designers envisaged. A similar attempt was made in the United Kingdom in the introduction of program analysis reviews (PAR), but again attempts to evaluate systematically the whole of government expenditure were unsuccessful. The degree of inertia in the system and the vested interests of existing institutions have proved too entrenched to be overcome by administrative procedure.

Full-employment budget. Although the idea of budget balance in the administrative budget has been the dominant consideration in the budgetary policy of most countries, it has gradually been realized that such a concept may be inappropriate when external shocks such as exchange rate movements or a world recession occur. Because varying levels of unemployment are a major reason why expenditures may change without comparable change in the public sector output, the concept of a full-employment budget has emerged. This type of budgeting is based on receipts and expenditures that would prevail under conditions of full employment. The approach views the actual expenditures and receipts for the coming year as of secondary importance; it assigns primary importance to the influence of the budget on the national economy. In time of recession a budget deficit may thus be presented as a necessary step toward achieving a balanced budget at full employment. Ideally, the budget should include estimates of expenditures and revenues at full employment, and also estimates of the same items at the anticipated level of employment. These ideas have been extensively used in the United States.

An analogous procedure could be used with respect to inflation, but this idea is still far from acceptance, because governments are no less reluctant to anticipate inflation than they are to budget for unemployment.

The U.S. full-employment budget was developed during World War II and has been regularly published in the president's annual Economic Report since 1962. Other countries have adopted similar measures as an aid to policy-making; for example, The Netherlands' "structural budget margin," developed in the early 1960s, and West Germany's "cyclically neutral budget," calculated by its German Council of Experts beginning in the late 1960s.

Adjustment for inflation

An analogous procedure could be used to correct for the impact of inflation. When inflation is rapid, interest rates are correspondingly high and a government may appear to run a deficit as a result of high debt servicing costs even if the real value of the debt is declining. The United Kingdom, for example, has seen a government deficit in almost every year in the postwar period, even though its debt has been a diminishing fraction of national income, because growth and inflation have been increasing the level of national income. Although inflation adjustments have been widely advocated and often adopted in private sector accounts, governments have been reluctant to adopt them for public finances because of the element of uncertainty in prediction.

Value for money measurements. As the emphasis in budgetary policy has shifted away from mere authorization of government spending and toward more public scrutiny of what government accomplishes, the idea of appraising value received for money spent in government finance has grown in importance. This has led to an increasing variety of measurements of public sector efficiency. In general terms, taxpayers need to be satisfied that their money is being used wisely. Because of the wide variety of items within even a single program, however, it is often difficult to identify precisely what is spent on the provision of each service, and the services that are provided rarely have well-developed private sector counterparts to act as a basis for comparison.

Obstacles to allocating efficiency

In some programs, governments have developed efficiency measures that relate observable facts, such as the quality of national health or the number of operations performed, to the cost of providing the service. The use of such measures is by no means widespread, however, and their basis is often open to question. The principal difficulty is that there is either no meaningful measure of the output of a public service—defense, for example—or output is complicated and multidimensional—as with education or health. The result is that any method used to measure efficiency is open to debate and challenge.

Attempts to control public expenditure, particularly since the mid-1970s, have led to some reexamination of which programs should remain in the public sector. In the United Kingdom many services (for example, hospital cleaning) have been transferred from public sector agencies to private contractors, in the search for more cost-efficient purchasing.

Budgetary planning: cash, volume, and cost terms. There are three principal bases for public expenditure planning: cash, volume, and cost. The cash basis is concerned simply with the projected money expenditure on the services involved. Making such projections is difficult because what the cash expenditure will buy depends on what happens to prices over the planning period. Moreover, many public expenditures cannot be planned in cash terms, because legislation prescribes the output. Most social benefits, for example, must be paid to anyone who is entitled to receive them, and this means that the government cannot control directly the amount of the expenditure.

The volume basis is concerned with the planned output of public services. The difficulties of measuring output, however, have already been noted. More often the planning process, assuming that changes in inputs are associated with changes in outputs, operates with reference to the cost basis of programs.

All countries have an annual program of public expenditure allocation, in which those responsible for individual programs argue for greater allocations for their activities and those responsible for raising the money attempt to control the amount allocated. In practice, the results of this process depend as much on the political weight of individuals in charge of a spending program as on an objective assessment of its desirability. The normal practice is to take as a base what each program spent the previous year and then argue about incremental changes, rather than (as under zero-base budgeting) to consider each program in its totality. This creates perverse incentives, in that departmental heads who have saved money in one area in a particular year have an incentive to spend more in other areas in order to protect next year's total budget.

The basis for most expenditure planning is therefore the number of public employees already in place and the volume of goods and services purchased in the base year. This, multiplied by base year prices, gives the input volume in the base year. In the late 20th century many countries (particularly the United Kingdom) have been abandoning this approach, largely because it gives inadequate control of total expenditure. One reason for a given volume's costing too much to supply is the so-called relative price effect. This arises because goods and services bought by the public sector (labour, medical care, or defense equipment) may rise in price more quickly than commodities generally. Once this has been determined, volume can be expressed in cost terms. The relative price effect is somewhat subjective, however, because of the difficulty of measuring the quality of goods and services. In the case particularly of health care and defense, the relative price effect will often contain the increased price of services and improved equipment, which are actually a volume increase.

The relative price effect

Cost measures, however, merely reflect the cost of a given input; controlling public expenditure in cost terms without taking full account of the relative price effect's change may lead to inappropriate volume responses or, more commonly, spiraling costs as existing input volume is maintained. Hence many countries have moved one stage further, attempting to monitor and control public expenditure in purely cash terms. The United Kingdom's public expenditure programs, for instance, are now "cash limited."

Although planning in cash has a superficial simplicity, at times of significant inflation it is not a very appropriate tool, and differential price rises may lead to a balance of expenditure provision somewhat different from the intended plan. In practice, although cash planning is presented as the base on which decisions are taken, those

countries that have adopted this approach in fact allow informal flexibility in cash budgets, with volume measures being implicitly, if not explicitly, adopted.

Components of the budget

In the United States the budget for each fiscal year contains detailed information on the outlays intended by the federal government and the receipts expected, including those from trust funds. The budget also divides authorized expenditure into that which can be carried out without action by Congress and that which requires further authorization. In any year, about half of federal expenditure requires authorization from Congress; by withholding this authorization, Congress is able to force changes in the government's budgetary policy. The budget also summarizes the outstanding debt of the federal government and estimates the size of the surplus or deficit expected on the basis of the revenue and expenditure projected in the budget.

Variations in U.S. and U.K. budgets

The U.S. budget is presented as a coherent whole for lengthy consideration by Congress, during which time it is often substantially revised. This joint consideration of revenue and expenditure is also common in most European countries. Practice in the United Kingdom, and in other countries with a British parliamentary tradition, continues to reflect the historical separation of revenue and expenditure. The U.K. budget consists of a number of different documents, with only limited attempts being made to relate one to another. A sketchy report of the government's intentions is given in an Autumn Statement, usually published in November, and detailed expenditure plans are provided in February or March in a White Paper. The U.K. budget, usually presented in March, is mainly concerned with taxation and is represented in a separate volume entitled the *Financial Statement and Budget Report.* This gives a general outline of budgetary strategy, details of proposed tax changes and estimates of likely revenues, as well as details of such items as capital receipts from asset sales and the size of the contingency reserve of unallocated money to cover unforeseen events.

Partly because of this fragmentation of the U.K. budget, and the difficulty of relating the public expenditure White Paper to the *Financial Statement and Budget Report,* debate is limited, and it is rare for any detail to be changed after the documents are published. The fragmentation of the budget is exacerbated further by the presentation of details of social security expenditure in yet another document.

EXPENDITURE

Composition of public expenditure. Expenditures authorized under a national budget are divided into two main categories. The first is the government purchase of goods and services in order to provide services such as education, health care, or defense. The second is the payment of social security and other transfers to individuals and the payment of subsidies to industrial and commercial companies. Both types are usually labeled "public expenditure," and in many countries attention usually focuses on the aggregate of the two. This obscures important differences in the economic significance of the two items, however. The first represents the public sector's claim on total national resources; the second the scale of its redistribution within the private sector.

The two main categories of public expenditure

In most Western countries, the share of the public sector in total economic activity averages between 20 and 30 percent. This reflects the proportion of workers who are employed in the public sector or in publicly financed activities, the proportion of national output generated there, and the proportion of incomes derived for productive services that is earned by public sector employees.

Some of these activities yield commercial revenues—the postal service, for example. Most have to be financed by taxation. In addition, the government raises taxation in order to redistribute income within the private sector of the economy. It taxes some activities and subsidizes others—through investment credits, for example. On a larger scale, it uses the benefit and social security system to make payments to needy individuals and raises taxes in order to subsidize those who warrant it. With this redistributive activity, plus the direct government productive activity financed from legislation, the total share of incomes taken in taxation is higher than the share of government in total production. It averages around 40 percent in Western economies.

In addition to direct expenditures, attention has been drawn to "tax expenditures." If the government favours a particular activity—such as investment—grants or tax concessions may be awarded to that activity. The two procedures have much the same effect on investment and on government revenues, but one appears to raise public expenditure and the other to reduce taxation. It has been suggested that these tax expenditures—tax reductions motivated by an economic or social objective—should be the subject of a tax expenditure budget similar to the public expenditure budget, and several countries have now moved in that direction.

For all private and public purposes within the economy, the scale of public activity is best measured as a proportion of national income: the total of incomes generated or (equivalently) of expenditures on goods and services.

The overall proportion of national income that is collected in taxes, raised from profits on government activities, or borrowed varies widely in the developed nations. This variation reflects different national decisions concerning the proportion of a nation's activity deemed most appropriate to have carried out by the various levels of government or by government agencies. Much of the variation occurs because of choices over the provision of health care (mostly public in the United Kingdom, mostly private in the United States) and over the level and importance of transfer payments.

By the late 20th century the share of national income devoted to public expenditure varied from almost 60 percent in countries such as Denmark, Sweden, and The Netherlands to about 30 percent in Australia, the United States, Japan, and Greece. The United Kingdom, Italy, France, and West Germany all devote between 40 and 50 percent of their national incomes to public spending.

Transfer expenditures

Expenditures on transfers also vary widely, depending partly on how redistributive the government wishes to be, partly on how much of this redistribution is carried out through the tax system, and partly on factors such as the number of old people and the level of unemployment. The dominant payment in every country is for old-age pensions, and the amount depends on how well developed private sector pensions are. Another factor is the extent to which the government chooses to use direct subsidies rather than tax concessions to stimulate the economy.

In the United States in the late 20th century, between 25 and 30 percent of the federal budget was being spent on defense and a similar amount on social security and Medicare payments. Only a fairly small proportion of the federal budget was spent on other items, with about 10 percent of the overall budget being devoted to the salaries and other remuneration of federal civilian employees. Most other provision of public services—education, roads, welfare, public health, hospitals, police, sanitation—were provided by state and local governments, which spent about three times as much as the federal government on the provision of civilian services. Both levels of government in the United States raise taxes from a variety of sources. The relative importance of state, local, and federal expenditure on civil functions has varied considerably, with the role of the federal government being greatest before World War II and declining after the war.

Public expenditure in Europe

In Europe public expenditure was both larger (as a share of national income) and more centralized during this same time. The United Kingdom, for example, devoted about 12 percent of national income to centrally funded social security programs; 5 percent to each of defense, the health service, and education; and smaller amounts to industrial support, law and order, and subsidies of various kinds. Although most revenue is raised centrally in the United Kingdom, administration of many programs is carried out at local levels, partly financed by a local property tax and partly through grants from the central government. Lo-

cal authorities are usually regarded as separate decision-making units, but the role of central government as a provider of finance that sets rules and imposes penalties has become dominant. (J.A.Ka./Ed.)

Growth of public expenditure. The proportion of national income devoted to public spending rose considerably during the 19th and 20th centuries. Much of this historical rise, however, cannot be taken as a direct measure of either the relative importance of government as a whole in economic decision making or of the comparative roles of central and lower levels of government. Inflation aside, in most countries the major reasons for the persistent rise in public spending since the middle of the 19th century have been war and the preparation for war, the rise in the cost of pensions for veterans, the great increase of the administrative role of government in response to expanded and urbanized populations, and the marked rise in the demand for a varied list of public services as the vote was gradually extended to the lower income classes.

Charles Bastable's observations

Writing in 1890, the Irish economist Charles Bastable observed that "in nearly all modern States outlay is steadily increasing," and "the older doctrines of economy and frugality have disappeared." He was referring to doctrines that had developed in the latter part of the 18th century, particularly in connection with the Industrial Revolution. He did not mean that there had been a "golden age" in which governments entirely refrained from interfering in the private sector. As Bastable himself pointed out, even the strictures of Anne-Robert-Jacques Turgot and Adam Smith on "excessive" government intervention did not preclude the encouragement of new industries.

In western Europe there was a long tradition of government influence on private economic decisions. The interventionist policies in the England of Henry VIII, Elizabeth I, and Oliver Cromwell, the France of Louis XIV and Colbert, and the Russia of Peter the Great are examples of such influence. But the sense of confidence conferred on the industrial class through the industrial and transportation revolutions of the 19th century, especially in Britain and the United States, produced an atmosphere that was unfavourable to government intervention. This did not, however, prevent rising pressure for government spending on economic resources, together with a secular rise in the magnitude and variety of the output of public goods that is still in evidence.

Expenditures after World War II

In the post-World War II period, government expenditures rose sharply. In the United States, overall public expenditure rose from 20 percent of gross domestic product to around 30 percent by the mid-1980s. Over this period, transfer payments as a proportion of national income nearly doubled. Other countries have seen an even steeper rise, both in expenditure on goods and services and in transfer payments. In Denmark public expenditure rose from about 18 percent of national income in 1950 to nearly 60 percent in the 1980s. The Netherlands experienced similar growth. (J.A.Ka./K.E.P.)

Problems of public expenditure control. The problems of controlling public expenditure vary across programs. Some are "demand led." Transfer payments, and particularly social security payments, are largely dependent on the number of old or unemployed people. Apart from reducing benefits (which may in turn be prevented by past commitments), or through macroeconomic policies designed to reduce unemployment, for example, there is little that can be done to limit these payments. Most countries have seen a steady rise in transfer payments as the longevity of the population and the benefits of pension schemes increase.

Public expenditure also depends on the price of the goods and services that the public sector buys and on the efficiency with which they are used. Public sector workers are often highly organized and may be well placed to demand pay increases from an employer who is able to recoup the costs from taxation. Public sector purchasing may be inefficient—civil servants may find it easier to enjoy a comfortable relationship with their suppliers, and, in fields such as health and military expenditures, administrators may demand the latest technologies with little regard for their cost-effectiveness.

At the same time, much of the public sector lacks the incentives to increase efficiency that apply to private firms in competitive markets. It is easier to resist innovation, and bureaucracies often have a conservative culture in which it is more important to avoid mistakes than to experiment with new techniques and procedures. With few external indicators of performance, managers in the public sector may feel inclined simply to promote the growth of their organization and the staff numbers and budgets that they control.

As the level and complexity of governmental involvement in the economy has risen, so public expenditure has become increasingly difficult to control. The only people with enough information to monitor their program needs are those actually engaged on the program. Coupled with technological change, the general tendency has been for expenditures to rise without any clear evidence of increased levels of service being provided. Indeed, in many key areas, such as health and education, expenditures have risen steadily at the same time that the public perceived a deterioration of service.

Resistance to rising expenditures

Governments in most countries have responded to this problem by occasional severe contraction of particular programs or of public expenditure in general. Numerous countries have adopted cost-cutting exercises with some limited success. But attempts at cost reduction can provoke inappropriate reactions. If politicians discover expenditure can be reduced without reducing the value of the services provided, they may insist on further cuts. If, on the other hand, popular or politically sensitive activities are restricted, there will be pressure to restore expenditures. Managers of public sector programs therefore often have incentives to respond to cuts in ways that maximize, rather than minimize, the effects on the services provided.

REVENUE

Governments acquire the resources to finance their expenditures through a number of different methods. In many cases, the most important of these by far is taxation. Governments, however, also have recourse to raising funds through the sale of their goods and services, and, because government budgets seldom balance, through borrowing. The subject of borrowing, because of the intricacies of deficit spending, is covered in a separate section of this article.

Taxation. Most countries raise resources through a variety of taxes, including direct taxes on wage and property income, contributions to trust funds, and a variety of indirect taxes on goods, either at the final point of sale or on the inputs used to make them. A smaller amount of revenue is raised from taxes on property, on capital gains, and on capital transfers, particularly at death. Most countries have a separate corporate income tax.

The composition of tax revenues. The balance between these different taxes has varied considerably over time and between countries. In the United States, sales taxes are relatively unimportant, accruing mainly to state and local governments. Federal government revenue is principally derived from taxes on personal and corporate income; until the 1980s the corporate share was diminishing, but changes in tax law tended to increase it. This dominant reliance on income taxes in the United States is a post-World War II phenomenon; at the beginning of the 20th century about half of all tax revenue came from taxes on property and half from sales taxes. Income tax was introduced on a regular basis only in 1913.

The tradition in Europe is somewhat different, with indirect taxes being relatively more important. All the countries in the European Communities impose a tax (at varying rates) on value added, charging tax on output from industry and rebating it on inputs. In the United Kingdom, value-added tax (VAT) raises about half as much as the personal income tax, and together excise duties and VAT raise about one-third of total tax revenue. U.K. corporation taxes on non-oil activities are relatively light, although oil revenues have become very important, despite fluctuations, contributing increasingly to all tax revenue.

Value-added tax

Australia, New Zealand, and the Scandinavian countries all rely heavily on income and profits taxes, which account

for about half of all revenue raised from taxation. In contrast, France, Greece, Portugal, and Spain only raise about one-fifth of their revenue from such taxes. Social security taxes are important throughout Europe, raising around 30 percent of all revenue in Austria, Belgium, France, West Germany, Greece, and Italy and rather more in The Netherlands. In contrast, the Scandinavian countries, Ireland, and the United Kingdom rely rather less on these taxes, which are not used at all in Australia and New Zealand. Japan, like the United States, raises about 30 percent of total tax revenue from social security taxes.

Payroll tax

Payroll taxes are relatively unimportant, raising significant amounts only in Australia, Austria, France, Ireland, and Sweden, but rarely exceeding 5 percent of total revenue. Property taxes rarely account for more than another 5 percent, with the United Kingdom being the exception in this case. Sales taxes, excise duties, and VAT account for nearly one-half of all revenue in Greece, Ireland, and Portugal, compared with less than one-fifth in Japan.

The relationship between tax rates and revenues. In deciding how to raise enough money to finance its expenditure program, a government faces a large number of different considerations. First, the tax system is complex, containing many different taxes, each often having a complex structure. Perhaps the major consideration is the effects on behaviour that particular tax rates will cause.

Income tax

Income tax has a graduated structure whereby no tax is paid on the first segment of income and then each subsequent segment is taxed at a higher rate than the previous one. In the United Kingdom most taxpayers pay tax at a uniform marginal rate, while other countries have more steeply rising rate schedules. Higher marginal tax rates make work less rewarding, which tends to reduce work effort. High marginal rates, however, may have less impact in some areas than others, a factor that needs to be considered when deciding who should bear the tax burden. Such considerations presumably have influenced the trend in many countries to tax the wealthiest groups.

Whatever the structure of the tax, the general proposition that increasing tax rates will reduce work effort usually holds; and this, in turn, tends to reduce tax revenue again. A vigorous debate has persisted over the "Laffer curve," which postulates that at some level of tax the disincentive effects will be so great as to mean that an increase in tax rates actually reduces revenue. This idea has been influential in leading governments to attempt to curtail the share of public expenditure in national income. The administration of Ronald W. Reagan in the United States cut taxes in 1981 in the hope of increasing revenue by stimulating the economy, and, while this succeeded to some extent, expenditures grew even more, causing a substantial increase in the budget deficit.

Effect of tax rates on consumption

Tax rates affect the pattern and level of consumption. Excise duties, value-added tax, and sales taxes all change the relative prices of goods and the attractiveness of consumption relative to saving. Once again, an increase in tax rates will generate responses that tend to cause a reduction in revenue, and, again, governments must balance the strength of these effects when deciding on which rates to increase. Other considerations, such as the protection of domestic industries, also affect such decisions.

Tax rates also affect commercial decisions, and the balance between individual and corporate taxes must reflect this. Accordingly, many countries have sought to attract new manufacturing industry with tax concessions. Finally, as rates rise, taxpayers seek more ways to avoid taxes. They employ tax advisers to find more tax-efficient routes, which, in particular, can involve a search for capital rather than income-yielding assets and the movement of activities overseas to less heavily taxed countries.

The balance between taxes. As the share of public expenditure in overall national income has risen, so has the strain on traditional sources of tax revenue. The original stalwarts, property and capital taxes, have shrunk in importance and been replaced by increasing reliance on income taxes, on social security contributions, and on sales taxes of various kinds. The balance between these taxes varies considerably among countries, which make differing decisions about the appropriate balance between taxes.

Each of the main types of tax is perceived by taxpayers in different ways. Social security taxes have everywhere risen in importance, partly as a result of the growth of social security expenditures but also because their association with the benefits received, however loose, reduces the unpopularity of increases. Income taxes tended to increase in many countries until the mid-1970s—even longer in the United States—because the exemptions and rate schedules were not fully indexed to inflation. This was later reversed, however, by indexation or explicit tax cuts.

Sales taxes are less obvious, as they change the price of goods a consumer buys rather than his income. At times of high inflation, it is often hard for taxpayers to identify what proportion of the price rise is actually caused by increased taxation, which has led to increasing reliance on this kind of tax. But sales taxes too have their limits; when the proportion of tax on a good is sufficiently high, consumption declines and there is political pressure from consumers and industry to reduce the tax increases. Governments have been reluctant to increase indirect taxes significantly as the control of inflation has become a major policy goal.

Sale of goods and services. Taxation is not the only means by which a government can raise revenue. It can charge for the services it provides, or it can undertake profitable commercial activities. This is done to some degree by all Western governments, although the revenue raised is much less than that raised by taxation.

Other sources of revenue

Charging for public services faces a number of difficulties. Perhaps the most important is collection costs. Public services such as roads and parks are difficult to charge for, because they are closely integrated with the community. Some countries have tolls on major highways and a few parks have admission charges, but in the main these are supplied free of charge. Other goods, such as museums and art galleries, are easier to charge for, but attempts at charging often generate more political opposition than can be justified by the limited revenue that could be raised.

A second consideration in deciding on charges is that it is rarely economically efficient to charge for public goods. Parks and roads, for example, have high initial costs but relatively low costs per user. Imposing a charge will mean that fewer people use them, and unless congestion is sufficiently severe to reduce others' enjoyment, less overall welfare will be generated.

In many countries, particularly the United Kingdom, many industries are publicly owned, and these include highly profitable industries such as those supplying gas or electricity. The profits from these industries provide revenue. It is an open question, however, whether they provide as much revenue as would the assets employed if these were invested in private sector companies. Other publicly owned undertakings make substantial losses, and this reduces the net value of any surplus. Few countries rely heavily on public sector industries for revenue, the nationalization of specific industries being usually justified on other grounds. (J.A.Ka.)

GOVERNMENT BORROWING

Although most of the resources required for public spending are raised each year through taxation, it is rare for any modern budget to balance in any one year. For a variety of reasons, ranging from a desire to accelerate capital spending to a policy of economic stabilization, governments may choose to raise some of their resources by borrowing rather than taxation. Most countries today run an annual budget deficit, and the deficits have tended to increase in size. For some countries—such as the United States and many developing countries—this means that the burden of the debt has been steadily increasing. Most European countries have a substantial accumulation of liabilities from wars and reconstruction, not shared by the United States. Even though they also run a current deficit, and their total debt may be increasing in size, the rate of increase is often much smaller than that in the United States and often below the rate of growth of national income. Thus, the burden of the debt is increasing less rapidly and, in some countries, may even fall. In times of inflation it may be possible for a government to run a

deficit without actually increasing the real burden of debt, as inflation erodes the real value of its existing debt.

Although most countries ran a deficit budget in the late 20th century in response to a world recession and high rates of inflation, only a minority did so in the 1960s; in the European Communities, on average over the period 1960–73, only Belgium, Ireland, Italy, The Netherlands, and the United Kingdom ran a deficit.

Forms of public debt. The necessity for governments to borrow in order to finance a deficit budget has led to the development of various forms of public debt, which are now a central feature of all capital markets. Governments may owe public debt in the form of bonds, notes, bills, and the like, which require specified payments to the holders at designated times. For the most part, public debt differs from private debt only in that it is an obligation of government rather than of private individuals or corporations.

Public versus private debt

Public debt may be classified according to various criteria.

External and internal debt. If the debt is held outside of the issuing jurisdiction, it is called external; if it is held within the jurisdiction, it is called internal. The U.S. national debt is almost entirely internal, while the debts of many developing countries and of local governments in the United States are largely external.

Maturity period. Public debt ranges in maturity downward from infinity to periods of a month or even a few days. Debt instruments without a maturity date, requiring merely the payment of interest, are often called consols. The name originated in Great Britain, where the first important indeterminate-period debt issue happened to be one that consolidated a number of separate issues.

A large portion of government debt consists of bonds with specific maturities of five years to 99 years or more. Twenty- and 30-year periods are common. These are often known as long-term or funded debt.

Debt of maturity less than five years is often called short-term or floating debt and may take several forms: notes, with maturities from one to five years; treasury bills, with maturities from one month to a year and often sold at auction; and certificates of indebtedness, with similar maturity periods but available at a fixed interest rate.

Debt liquidity

The length of the maturity period affects what is known as the liquidity of the debt—*i.e.,* how quickly it can be converted into money. Securities with very short maturity periods are constantly repayable in money and thus have maximum liquidity. As the period of maturity increases, the liquidity falls, unless a capital loss is to be incurred, and the pure debt characteristic increases.

Type of issuer. Government debt may be directly issued by a government or by semiautonomous governmental organizations. Examples of the latter would include the railways and provincial power authorities in Canada and various federal lending agencies in the United States. Their issues may be guaranteed by the government (general obligation bonds) or may rest solely upon the enterprises themselves, to be paid out of their revenues. In the United States the latter type of obligation is known as a revenue bond.

Marketability. The great bulk of all government debt consists of marketable securities. These securities are negotiable and are sold freely on the market. They are usually issued in relatively large denominations, $1,000 or higher, and interest is paid by check or coupon on a periodic basis. Since they are salable, their price fluctuates from time to time, going above maturity value when the current market interest rate falls below the interest rate that they bear and falling below the maturity value when the current rate rises or when fear about the ability of the government to pay interest develops.

Other bonds bought by the public are not marketable but can be redeemed, at least after a specified period, for their principle plus accrued interest. Various savings bonds, including those of the United States, are of this kind.

Other characteristics. Bondholders may receive current interest either by redemption of coupons attached to the bonds or by check from the government. Alternatively, interest may be receivable only upon maturity or redemption of the bond, as in the case of savings bonds. Interest and principal are usually payable in fixed monetary units, but they may be payable in amounts with fixed purchasing power based on changes in price levels.

Economics of government borrowing. Government borrowing is likely to have effects upon the economy substantially different from those of other methods of financing, and the existence of a sizable debt may likewise have important consequences. The effects of retiring (or repaying) the debt may also be significant. National government borrowing has the greatest impact, but that of subordinate units may have some influence as well.

Effects of borrowing. Government borrowing in the strict sense includes only borrowing from the private sector of the economy—from individuals, corporations, and various financial institutions, including banks. When the government obtains its funds from the central bank (the Bank of England, the Bank of Italy, the Bank of Japan, or the Federal Reserve System in the United States), it is really creating money rather than borrowing it, since the purchasing power is made by the central bank and no obligations to the public are created.

How governments borrow

When a government borrows, funds are transferred from the lender to the government, the lender exchanging his money for government securities. The effect is to reduce the liquidity of the lender—his command over cash—to an extent dependent upon the nature of the securities. The reduction in liquidity is small with short-term securities and greatest with nonsalable, nonredeemable securities—a type seldom issued except in time of war or other crises that create financial emergencies.

Funds loaned to the government almost certainly come from savings, unlike, for example, funds paid in higher taxes, which are more likely to come out of consumption. In many countries the major holders of public debt are, in fact, pension funds, which invest in government debt on behalf of the individual members of their pension schemes. To pay higher taxes, many individuals are forced to reduce their consumption since they have no margin of savings and are unable or unwilling to go into debt; others do so as a matter of choice, in an effort to keep their savings intact. Lending, on the other hand, is entirely voluntary. The person who buys government securities is not likely to increase his rate of saving or to decrease his consumption. If government borrowing raises the market rate of interest, this may in turn encourage the diversion of additional money to saving, as may government securities that offer additional attractions—such as small denominations or redeemability—not possessed by other securities. But both effects in total are not likely to be of any particular significance.

The net effect of government borrowing on total spending and thus on employment and national income depends upon its influence on real investment—the purchase of new capital goods. In a period of unemployment, when savings are available in greater quantity than is required for investment, government borrowing does not compete with private investment nor make it more costly. In effect, the government absorbs funds that would otherwise be idle.

Borrowing during full employment

In periods of full employment the situation is substantially different. With banks loaned up to the limit of their reserves and real investment absorbing all of savings, government borrowing will restrict private spending as much as an increase in taxation will under the same conditions.

Government borrowing is of economic significance in several other respects. First, the buying and selling of government securities provides the central bank with a means of influencing the money supply, essential for effective monetary policy. Second, borrowing avoids the adverse effects that taxes may have on incentives, particularly if the taxes are raised sharply above levels to which persons have become accustomed. Third, borrowing permits government expenditures to be higher than would otherwise be feasible. Finally, the foreign borrowing of some governments gives them access to a greater quantity of foreign exchange, which enables them to finance the import of capital goods essential for economic growth. This consideration is not of concern to highly developed countries.

Effects of debt. The existence of a government debt is of

economic significance in itself, as distinct from the effects of the borrowing. In the first place, individuals who hold government securities regard them as a portion of their personal wealth. This is true even though the only way the government will ever pay the interest on the debt or repay the principal is by levying taxes on the community, which holds the debt. In this sense "we owe it to ourselves." But since these links are not immediately apparent, the existence of a debt may make individuals spend more on consumption and save less than they otherwise would. The additional consumption may reduce the rate of capital formation and economic growth; it may also increase the level of employment over what it would otherwise be.

Second, because government securities are more liquid than most other investments, their holders are able to increase consumption out of accumulated savings more easily than they could otherwise. This may contribute to inflationary pressures.

Third, if investors, and particularly the business community, regard the national debt as a source of potential economic instability, their willingness to undertake real investment will be lessened. At times, particularly in the 1930s, there has been widespread fear of government debt even though there was, in reality, little basis for the fear. A similar phenomenon sometimes arises in the case of subordinate units of government. A large debt may discourage expansion of economic activity because of the fear of high taxes in the future and the realization that the large debt may prevent borrowing for urgently needed local improvements.

Resistance to national debt

When governments borrow they must meet interest obligations, and these are usually paid out of taxes. The payment of interest on government debt thus involves a transfer of wealth from taxpayers to bondholders. The taxes may have adverse effects upon incentives, while receipt of the interest will provide no offset to these adverse effects. The tax-and-interest-payment program is also likely to redistribute wealth in favour of higher income groups, since government bonds are likely to be held to a greater extent by those groups. The effect may be to increase saving and reduce consumption.

Finally, large interest obligations lessen the ability of the government to finance other governmental activities. This effect is particularly obvious at the local level, where there are limited tax potentials.

Retiring the debt. The retirement of government debt arising from a budget surplus has effects opposite from those of borrowing. Bondholders receive money in exchange for their bonds; though they could increase their consumption, they are more likely to put the funds into other securities and, as a consequence, security prices rise and money capital becomes more readily available for business investment. Whether it is used for that purpose depends, of course, on factors within the existing general economic situation.

Money for retirement must be obtained from some source. If it is simply created, there is no repressive effect on consumption or investment, and total spending in the economy rises—although by an amount relatively small compared to the total retirement. If, as is more common, the debt is retired from tax revenues, consumption is reduced in substantial measure; the remainder of the tax is absorbed from savings, and real investment may be reduced. The net curtailment in spending from the program of debt retirement is likely to reduce total spending in the economy. Elimination of the debt has one other effect: while current taxes will be increased, future taxes required to meet interest and principal obligations will be reduced.

Use of taxes for retiring debt

It is commonly thought that borrowing shifts the burden of governmental activities to future generations, since those generations will be assessed higher taxes to pay the interest and principal. Some economists have disputed the idea, noting that future generations will inherit both the bonds and the obligations to pay them and collectively will be neither richer nor poorer than if the debt had not been incurred, except as a result of the difficulties incident to the debt and its retirement noted in preceding sections. Regardless of the methods of financing, the real cost of any governmental activity, war or otherwise, is borne in the form of reduced private consumption and investment and harder work or the like during the period in which it is carried on. The only burden on the future is that arising from the depletion of natural resources, and this is not affected by methods of financing. Nevertheless, the method of financing may affect the way in which the burden of public expenditure is shared among different groups, including age groups.

Limitations on public sector debt. Although borrowing can often seem an attractive alternative to raising money from taxation or indeed to spending less, there are limits to how far a government can allow itself to become in debt either to its own citizens or to overseas investors (including intergovernmental agencies, such as the International Monetary Fund). In many countries, particularly in Latin America and in Africa, these limits have been exceeded in the late 20th century, with serious results for the stability of the country concerned. When debt rises to unacceptable levels, so that investors cease to believe in the ability of the country's tax base to support it, then drastic measures are forced upon the country, including severe contraction of the economy.

Problems of borrowing. The desirability of government borrowing has been debated for centuries. The traditional argument against borrowing is, of course, the interest burden to which it gives rise, an argument applicable equally to private and governmental borrowing. These interest obligations require either higher levels of taxes, with possibly adverse effects on the economy, or reduced expenditures for other purposes. The payment of interest may easily result in a transfer of purchasing power to higher income groups, contrary to accepted standards of equity.

The interest burden

As well as causing more and more of the government's resources to be used to pay interest on its debt, a large public debt can push interest rates up for other borrowers. If the government is persuading a high proportion of available funds to be spent on public debt, the amount remaining for investment in other places, for example, investment in industry, is correspondingly small, with the result that a higher price (or higher interest rate) needs to be paid to attract such investment. This has been seen as a serious constraint in recent years, and most Western governments have tried to reduce their borrowing in order to keep interest rates down. There is some debate over just how important public borrowing is for interest rates, with monetarists believing it to be extremely important. Other types of economists are more skeptical, contending that factors such as inflation and the availability of private sector investment opportunities are more significant.

The financing of expenditures by borrowing instead of taxation and the debt itself, once incurred, increase total spending and so tend to produce higher prices and other inflationary effects in periods of full employment. During periods of full employment, any increase in government expenditures not offset by an equivalent decline in private spending for consumption or business expansion will be inflationary. This is the usual argument made against the use of borrowing instead of taxation from the standpoint of the goal of economic stability. It is primarily relevant to national government borrowing because the national government must assume the primary responsibility for lessening economic instability. But state and local borrowing is, of course, equally inflationary.

Borrowing, if freely employed, can easily lead to increases in government expenditures beyond levels regarded by society as the optimum and may reduce the pressures for efficiency and elimination of waste. As governments consider expenditure levels, the adverse reaction to taxation serves as an offset against the favourable response to increased services that will have to be paid for by taxation and thus facilitates the attainment of a balance between government-produced services and privately produced services. But if borrowing replaces taxation and is generally accepted as a suitable routine method of financing, the pendulum will swing in the direction of increased governmental activity, and appropriate balancing will be lost. The best evidence of this danger is to be found in the history of state and local government finance in the early 19th century in the United States, when large sums of

money were borrowed for purposes of limited usefulness to society. Borrowing appears to be a less painful method of financing government, but, as has been noted, the costs of public expenditure still have to be met from current consumption or investment.

Restrictions on borrowing. Efforts have been made in some countries to set restrictions on government borrowing through legislative acts. In the United States, fear of excessive borrowing has resulted in restrictions on the amounts the executive, and even the legislative branches of government, can borrow. When many states found themselves in financial difficulties after borrowing heavily to provide funds for canals and railroads in the middle of the 19th century, public debt provisions were written into the constitutions of all but seven states. The provisions limiting borrowing differ widely. In most jurisdictions a maximum, usually expressed as an absolute dollar sum and one relatively low in terms of present-day expenditure levels, is set. Either this figure cannot be exceeded at all (except by amending the constitution) or it can be exceeded only with the approval of the voters at an election. In some places all bond issues require approval by popular vote and in some instances by more than a bare majority. The purposes for which funds may be borrowed and the duration of the issue are also frequently restricted. These constitutional restrictions have unquestionably lessened state borrowing; in so doing they have, perhaps, reduced waste, but they have also sometimes prevented urgently sought improvements. The limits have likewise greatly increased the use of revenue bonds, which are normally not subject to the restrictions. Unfortunately, the interest rate on these bonds is higher than the rate on other bonds.

Variations on limiting borrowing

U.S. state restrictions on borrowing

Restrictions on municipal borrowing in the United States are almost universal. The restrictions, established either in the state constitutions or by state legislation, limit the total sum to be borrowed by any particular unit to a certain percentage (from 2 percent to more than 20 percent) of the total assessed value of its property. The limits vary for different types of local units (city, county, school district, etc.). They usually do not apply to debts incurred for self-liquidating enterprises. In many states every bond issue must be approved by popular vote, in some instances by a two-thirds majority. In other states the limits established may be exceeded by popular vote, often with a requirement beyond a mere majority. Legislative controls also include maximum interest rates that may be paid, the duration of the issues, the purposes of the borrowing, and the establishment of means of retiring the bonds. Several states exercise review over local bond issues. Like the states, the local governments have found means of escaping the restrictions. Special taxing districts with their own debt limits are often formed when a city has reached its limit. Revenue bonds are also employed. In some states, such as Pennsylvania, there has been widespread creation of special authorities, school building authorities, for example, that have been established with the power to finance the building of schools by issuing revenue bonds. In turn, the authority pays interest and principal on the bonds from rentals obtained from the school districts using the buildings.

U.S. Congressional restrictions on borrowing

While there are no constitutional limits on federal borrowing powers in the United States, Congress for many years has restricted borrowing by the Treasury Department. Before 1917 borrowing was permitted only upon specific authorization by Congress. After 1917 maximum figures were set at first for each type of loan and then, after 1938, as an overall total. The 1938 figure of $45,000,000,000 was gradually increased to a high of $300,000,000,000 in 1945 and reduced to $275,000,000,000 in 1946. Buttressed by a strong belief prevailing in Congress that refusal to raise the limit would check growth in government spending, the limit remained at the 1946 level until 1954. Eventually, pressure on the limit became so great that various government bodies such as lending agencies were forced to borrow on their own at higher interest rates. A series of increases was made in the 1960s and 1970s, and by the early 1980s the limit exceeded $1,000,000,000,000. Experts differ in their estimates of the usefulness of the federal limit. Some believe that it curtails government waste and unjustified increases in expenditures, while others argue that it reduces flexibility in meeting emergencies, checks needed increases in various activities, could prevent quick action to stave off a depression, and leads to uneconomical forms of borrowing.

By the mid-1980s, the U.S. deficit approached an annual figure of $200,000,000,000 and was seen as a central economic problem. A movement grew for a constitutional amendment to prescribe a balanced federal budget. Such a constitutional provision would not, however, specify how such an outcome was to be achieved. Nor, given the many budgetary concepts described, would balance easily be defined. Congress instead passed the Gramm-Rudman-Hollings Act in 1985, which required arbitrary reductions in spending in all programs if the overall deficit failed to fall within certain limits that were set for the purpose of eliminating the deficit by the end of the decade.

Borrowing restrictions in Canada

In Canada, neither the dominion nor provincial governments are subject to debt limitations. Local government limits are comparable to those in the United States, and in several provinces bond issues must receive the approval of a provincial agency. In the United Kingdom borrowing by local governments is subject to control, and limits are usually established in terms of a ratio of debt to total ratable value (assessed value of property). After World War II much local borrowing was channeled through the Public Works Loan Board, and thus was subject to additional control. There are no arbitrary limitations on the amount the U.K. central government may borrow; effective limits are set by the reaction of capital markets and of interest rates to borrowing.

Wartime finance. The use of borrowing is regarded as inevitable in periods of major war. If taxes were increased sufficiently to finance all war costs, they could seriously impede the war effort by impairing incentives to work and by reducing the overall morale of the people. The limits of economically and politically tolerable taxation may well be below the maximum feasible allocation of resources to the war effort. Adequate tax increases would also aggravate the inequities of the tax structure; an overall level that would reduce total consumer spending to a level equal to the rate of output of consumer goods might well push some persons below subsistence levels and make it impossible for others to meet fixed commitments. While the use of borrowing as a method of war finance makes the control of inflation more difficult, there appears to be no escape from the necessity.

Evolution of government borrowing. The evolution of government borrowing was very slow. The extensive use of loans by governments became possible only after the ruler had become differentiated from the state and after the fact of the continuity of the state had been separated from the persons of the rulers. Other factors were also required: the development of a regular revenue source to provide funds for repayment of loans, a monetary system, and an organized money market. The earliest loans of medieval times were either forced loans or personal borrowing by the sovereign. Government borrowing in its modern form first occurred in medieval Genoa and Venice when the city governments borrowed on a commercial basis from the newly developed banks.

Early forms of government borrowing

France. Throughout much of French history public borrowing has been of major dimensions. Ministers of finance in the 17th and 18th centuries found the problem of managing the debt almost insuperable. During the Revolution that began in 1789, about two-thirds of the accumulated debt was repudiated, and the remainder was refunded in new securities in 1800 at a total of 926,000,000 francs. The sum increased by only 340,000,000 francs during the Napoleonic period because Napoleon's military expenditures were financed mainly by foreign levies. A large increase occurred during the Second Empire, when the debt rose from 5,516,000,000 francs in 1852 to 12,310,000,000 francs in 1870. The Franco-German War, which ended in defeat for France, and the consequent imposition of an indemnity of 5,000,000,000 francs by the victorious Germans raised the French public debt to more than 21,000,000,000 francs in 1873. Most of the increase was financed by four bond issues. After 1878 the

Loss of the franc's value

debt increased further as a result of public works expenditures and France's colonial expansion until it stood at 34,204,000,000 francs at the outbreak of World War I. The war and its effects multiplied the debt, although at the same time inflation reduced the value of the franc by half. The inflationary trend continued throughout the interwar years, and by 1960 the franc had lost more than 99 percent of its 1914 value. The increase of the public debt in this period to 8,404,000,000,000 francs has to be seen, therefore, in the context of the continuing inflation. The issuance in 1960 of a new franc equaling 100 old francs automatically reduced the nominal value of the public debt to 1 percent of its previous figure. Following the introduction of the new franc, the national debt continued to rise.

United Kingdom. Government borrowing in the United Kingdom dates to the end of the 17th century. In 1692 legislation pledged the receipts from beer and liquor taxes as security for a loan of £1,000,000. The trend of the debt was upward throughout the next 150 years largely because of wars; by 1802 it had reached £523,000,000 and by 1840, £827,000,000. The second half of the 19th century saw gradual reduction of the debt to £610,000,000 in 1900, while the amount of debt still remaining became less significant because of the growth of the economy in the same period. World War I brought a tremendous increase, the 1920 figure being £7,828,000,000. The 1920s showed little reduction, and the figure rose slightly during the Depression years. World War II brought the level to £21,366,000,000 in 1945, and the figure rose in the postwar period—partly as a result of nationalization of industry. During the 1980s it surpassed £140,000,000,000.

United States. In the United States, when the federal government was formed, it assumed the debts of the states and various other obligations incurred during the American Revolution, all of which were funded into a single debt issue of $75,000,000 in 1790. The government was highly successful in avoiding additional borrowing in the early years, except for the War of 1812, and during 1835 all federal debt was eliminated. The years 1835 and 1836 were the only ones in the history of the country during which there was no federal debt at all. The American Civil War, only 25 percent of which was financed by taxation, pushed the debt to a total of $2,678,000,000 in 1865. Most of this debt was retired by budget surpluses during the following decades; debt reduction proceeded so far that bonds available for security behind national bank notes became inadequate. The debt remained relatively constant in the 1890s and during the early 1900s. World War I brought an increase to $26,000,000,000, consisting in part of short-term and intermediate-term securities and in part of Liberty Loan bonds. In the 1920s the government was able to reduce the debt; the low point reached was $16,185,000,000 in 1930, primarily by budget surpluses.

The 1930s brought budget deficits because of the Depression and the efforts to stimulate recovery. Despite extensive borrowing, which raised the total debt to $42,968,000,000 by 1940, interest rates fell sharply as a result of the surplus of money capital and federal reserve action. A substantial part of the borrowing was on a short-term basis, partly because the interest on such loans was extremely low. With the outbreak of World War II, borrowing rose sharply and by 1946 the debt had reached $269,000,000,000.

U.S. borrowing during World War II

In the postwar period the debt fell to a low of $252,000,000,000 in 1948, then gradually rose. This increase was caused by budget deficits arising primarily from a high level of defense spending and the unwillingness of Congress to hold taxes to rates high enough to meet the expense and in some years from a desire to stimulate economic activity. During the 1970s the debt increased each fiscal year; by the mid-1980s it had passed $1,400,000,000,000, and it continued to grow, although some factions sought legislation that would put a ceiling on the national debt.

U.S. state government borrowing

The states incurred substantial debts in the early part of the 19th century, largely for public improvements, and some found themselves in financial difficulties. As a result, borrowing came nearly to an end until after 1900; after that date there was further borrowing, particularly for highways. After 1945 the state debt increased sharply and had passed $167,000,000,000 by the mid-1980s. Much of this additional borrowing was for highway purposes. The local governments have traditionally borrowed more than the states, largely because of the nature of their functions. Local debt in the 20th century increased steadily and had passed $287,000,000,000 by the mid-1980s.

Canada. Canada's debt began with $75,000,000 (Canadian) at the time of confederation in 1867, when certain obligations were taken over from the provinces. The figure grew slowly until 1915, largely because of government railroad financing. World War I pushed the figure to $3,042,000,000 by 1920; the total rose as the Canadian National Railway was developed, fell slightly in the late 1920s, rose to $5,000,000,000 with Depression borrowing, and reached $15,713,000,000 at the end of World War II. Some debt fluctuation then took place and the figure reached about $17,000,000,000 by 1950. By April 1969 it had risen to $35,800,000,000 as a result of deficits. Canadian debt continued to rise until 1976, when it briefly decreased by about 5 percent from the previous year. By the mid-1980s the country's debt had surpassed $160,000,000,000. The path of provincial and local borrowing in Canada was similar to that in the United States, though with a slower rate of growth.

Germany. The German Reich, founded in 1871, began as a confederation of sovereign states. Most financial powers remained with the individual states until the Weimar Republic was established in 1919. A French war indemnity of 1871 was used largely to reduce the public debts of the states. As late as 1913 the debt of the Reich (4,900,000,000 marks) was less than half that of Prussia (9,900,000,000 marks) and substantially less than the aggregate debt of all the other federal states (6,300,000,000 marks). The country's defeat in World War I led to financial chaos. In 1925, after the stabilization of the new Reichsmark, the public debt was 2,413,000,000 marks. In the 1930s the public debt rose, going to 52,060,000,000 marks by 1940. World War II was financed mainly by borrowing, from both the private sector and the central bank; by 1945 the debt stood at more than 300,000,000,000 marks. Most of this was wiped out by the postwar currency reform of 1948. Following this currency reform, West German public debt increased nearly fourfold in the 1950s, twofold in the 1960s, and fivefold in the 1970s; by the mid-1980s it had surpassed 360,000,000,000 marks.

German public debt after World War I

Japan. The rise of the modern Japanese state began in the latter part of the 19th century. The government began to issue bonds in 1870. The cost of financing the war with China in 1894–95 and a subsequent buildup of its army and navy raised Japan's public debt from 255,000,000 yen in 1890 to 506,000,000 in 1900. The war with Russia in 1904–05 cost about 1,500,000,000 yen, which was mainly raised by foreign borrowing. The financial burden of the growing empire was henceforth largely covered by taxation, so that public debt did not increase substantially from 1907 until the end of World War I. Between 1918 and 1930, however, the debt doubled. In these years a large proportion of the debt was in foreign-owned bonds. In the 1930s the government adopted heavy spending policies, mainly for military purposes, and in 1940 the debt was more than three times what it had been in 1930. Between World War II and the mid-1980s the debt had risen from 150,795,000,000 yen to more than 111,000,000,000,000.

Local governments in Japan have always been heavy borrowers. This has continued to be true in the postwar years, when prefectures, cities, towns, and villages issued bonds on a scale approaching that of the national debt. Much of the local indebtedness was used to finance large public works programs.

Debt and national income. The absolute figures of growth in government debt exaggerate the actual growth in the debt relative to the economy as a whole. In the first place, the general price level has increased significantly over recent decades; since debt obligations are stated in fixed monetary terms, the relative magnitude goes down as the price level goes up. The general rise in prices over a period thus reduces the problems created by the debt for the government and the magnitude of the adverse effects

of the interest payments on the economy. The gain occurs at the expense of the bondholders, whose real economic position is worsened by the change.

Second, the rise in national income reflecting an increase in output reduces the real significance of a fixed sum of debt for the economy. The combined effects of the real and monetary influences can be illustrated by expressing the size of the debt as a ratio to gross national product (GNP) over a period of years. In the United States the ratio fell from 129 percent in 1946 to 35 percent in 1980. It had risen again slightly by the mid-1980s. The ratio of interest payments to national income likewise fell until 1968, when it began to increase, reaching 3.8 percent in 1980. In the United Kingdom the ratio of national debt to GNP fell from 221 percent in 1952 to 136 percent in 1958. The ratio continued declining to less than 100 percent in the mid-1960s and less than 50 percent in the mid-1970s, although the size of the debt increased slightly over the period. By the early 1980s the ratio of national debt to GNP was about 43 percent.

Relating national debt to GNP

Comparing debt in various countries. An adequate comparison of debt burdens in various countries is difficult to make. The reported figures are by no means entirely comparable because they vary in their treatment of debt incurred for various commercial enterprises, loans from foreign countries, special issues, and the like. The relative importance of the national debt and the debt of subordinate units of government also varies, and figures for the latter are not available for many countries. Any comparison of absolute figures of debt in monetary terms is of limited value and may be very misleading because of problems of conversion to a common monetary unit. The only meaningful figure is the ratio of national debt to national income, and the significance of these figures is greatly lessened by the inaccuracy of national income data for many countries.

Sovereign debt. The oil crisis of 1973–74 and its aftermath created a new instability in world capital markets. Some countries, particularly Middle East producers with few economic activities not based on oil, gained revenues much in excess of their capacity to spend. Others, particularly in the less developed world, faced balance-of-payments problems that they found difficult to cover. Some other oil producers, such as Mexico, borrowed heavily in anticipation of rapidly increasing revenues. Those countries with surpluses of revenues over expenditures wanted to retain the liquidity of the financial assets that they acquired, and Western banks increasingly took on the role of intermediaries between the surplus and deficit countries. This led to the growth of sovereign lending—bank lending either to governments or to agencies of governments with government guarantees. While a bank lending to a private individual or company normally requires examination of the relationship of the loan to the borrower's assets, and of the interest to income or cash flow, banks felt able to apply more relaxed criteria to sovereign loans.

By the early 1980s, however, it was apparent that for many countries sovereign debt had grown to levels at which even the interest on these loans would be met only by further borrowings. Moreover, these countries' limited capacity to repay might be undermined by political or economic instability. The problem was particularly acute in Latin America, where U.S. banks had lent aggressively. Argentina, Brazil, and Mexico had very large external debts; smaller countries such as Bolivia, Ecuador, and Peru had debt burdens that were even larger in relation to their capacity to service them. Similar difficulties were encountered in Africa and in parts of eastern Europe, particularly Poland.

Complications of sovereign borrowing

The debtor countries were reluctant to repudiate their debts, which would have deprived them of access to the world capital markets and even perhaps to the world trading and payments system for a considerable time. At the same time, the lending banks were reluctant to demand repayment of their loans, which would have led to default and losses that would have wiped out a substantial portion of their reserves. Thus, there was a mutual interest in using the financial system to continue to support the indebted governments, and, paradoxically, the negotiating position of the borrowers was stronger than that of the lenders.

These were highly unstable arrangements, arousing fears that major defaults would occur. Such defaults might well set off a cumulative process of demands for repayment and defaults that would undermine not only the economies of the debtor countries but also the banking and financial systems of the countries in which the lending institutions were located. Avoiding such a crisis has demanded continued sensitive responses on the part of international financial agencies, such as the International Monetary Fund and national regulatory institutions.

(J.F.D./J.A.Ka.)

The budgetary process

The budgetary process is the means by which the executive and legislative branches together formulate a coherent set of taxing and spending proposals. The mechanics of this process, and the relative roles of the two parts of government, differ considerably among countries.

GOVERNMENT VERSUS PRIVATE SECTOR BUDGETING

Although the process of preparing and discussing a national budget has progressed considerably during the 20th century, it is in a number of senses still inferior to the way budgeting is carried out by private sector companies or indeed by individuals. Commercial practice is governed by a series of well-defined rules, and firms are required to produce a balance sheet, a profit and loss account, and to monitor their cash flow carefully. The total indebtedness of a company is monitored closely by its shareholders, who are also critical of future forecasts of profits and growth. Individuals who fail to budget adequately are equally closely monitored by bank managers and credit agencies, and those with complicated affairs can draw upon skilled professional help.

The accountability of government, even in a well-developed democracy, is in reality considerably less acute, or certainly less clear, than that of companies to their shareholders or individuals to their various creditors. As a result, governmental budgeting is frequently of lower quality than is the norm in the private sector. Forecasts of receipts and expenditures are often wildly at variance with reality; changes to accounting practices are sometimes made for cosmetic political purposes; and certain distinctions, such as those between capital and current expenditures, are frequently blurred deliberately.

Problems of government accountability

These criticisms of the national budgetary process are more valid in some countries than in others. The extent of scrutiny of the national budget varies widely, and governments vary in how ready they are to provide relevant information and to what degree they try to obscure features of the budget by complicated and disjointed presentation. The United States has a relatively open budget, which is presented as a whole and subjected to congressional scrutiny. In contrast, the government of the United Kingdom presents the budget in different documents at different times, and, although subject to parliamentary scrutiny, it is rarely changed.

SELECTED NATIONAL BUDGETARY PROCEDURES

The United States. Since 1921 the budget of the United States has been the responsibility of the president. It is prepared under his direct authority by the Office of Management and Budget (OMB). The process begins when the various departments and agencies prepare their appropriation requests, based on expenditures required under existing law and those estimated under new legislation to be proposed by the president. These requests are carefully scrutinized by the OMB. In case of disagreement, Cabinet officers negotiate directly with the president, who is ultimately responsible.

Unlike the budgets of many countries, that of the United States deals mainly with expenditures. Revenue is covered in much less detail. Great significance is ascribed to the size of the expected deficit or surplus, even though there is no legal requirement that the budget be in balance.

The budget is submitted in January and normally ap-

plies to appropriations for the fiscal year beginning July 1. These must normally be spent in the following two years. For some items, such as construction or procurement of military hardware, appropriations are made to cover expenditures for the whole construction period.

The Appropriations Committees of Congress

When the budget reaches the House of Representatives, it is distributed among the subcommittees of the Appropriations Committee. Each subcommittee is concerned with a particular organizational unit. There is virtually no consideration of the budget as a whole by the committee as a whole. Revenues fall under the jurisdiction of the Ways and Means Committee of the House and are considered separately and possibly even at a different time from appropriations. The upper house of Congress, the Senate, plays a secondary role with respect to the budget. Its Appropriations Committee acts as a kind of court of appeal from the House Appropriations Committee. These procedures allow more coordination than appearance would suggest. The committee chairmen are among the most influential members of Congress, and the committee staffs are experienced and skillful.

Economic impact. The president sends three documents to Congress in January: the State of the Union Message, the Budget, and the Economic Report. The first is addressed to broad national policy, whereas the Economic Report is concerned with economic policy alone. In particular, it seeks to assess the economic impact of the budget and its effect on employment and prices. It is therefore mainly concerned with the stabilization rather than the administrative aspects of the budget.

Balance-of-payments considerations

Budget makers in the United States must also consider the international balance of payments. This is a relatively new problem for the United States, which historically enjoyed a balance-of-payments surplus until the 1960s. Since that time declining trade balance, coupled with heavy military commitments abroad, has required that budgetary and other economic measures be designed with international as well as national economic balance in mind.

Effect on resource allocation. There is no formal machinery for ensuring that the budget strikes a satisfactory balance among the different programs contained in it. But the alternatives receive a good deal of scrutiny. Cabinet officers have their clienteles in the public and in Congress and through them can bring political power to bear in the competition for funds. Similarly, the congressional committees are able to exert some influence on the budget during its preparation. The president for his part is not passive; he has a political position of his own that permits him to assert his conception of the national interest. Other pressures come from the taxpayers, who are able to express their stand on expenditures both to Congress and to the president.

A satisfactory outcome for the allocative process depends on the evenness with which these competing forces are balanced. Opinions differ. It can be said that the process of decision making is at least an open one, even though vested interests have entrenched themselves in particular areas. Although amendments to the budget are usually minor, the influence of Congress is not negligible. The prospect of facing Congress is a sobering one to the officers of government; congressional committees often strongly influence departmental budget recommendations. The secretary of agriculture, for example, may have a relationship with the agricultural committees of Congress that is closer than his relationship with the president.

The United Kingdom. The U.K. budget is submitted to Parliament by the chancellor of the Exchequer, who is responsible for its preparation. The emphasis of the chancellor's budget speech is on taxation and the state of the economy, rather than on the detail of expenditures; public discussion is devoted mainly to the chancellor's tax proposals. The estimates of expenditures are sent to Parliament with less fanfare and are reviewed by the departmental select committees on estimates of the House of Commons. The reviews hardly touch matters of policy. While the committees do not amend the budget, they are influential through their criticism and advice.

Authority of the Treasury in Britain

In the preparation of the budget, the Treasury appears to have virtually complete authority over the government departments on matters of detail. Major issues are settled in Cabinet discussions, the records of which are not available. The British system thus vests extensive controls in the Treasury bureaucracy.

A major part of the budget speech by the chancellor of the Exchequer is addressed to forecasts of employment, prices, and the balance of payments, together with a discussion of fiscal and monetary policies. Economic analysis is a continuing preoccupation of the Treasury. Forecasts are prepared three times a year, although only published once a year with little detail; a budget committee composed of important financial and economic officials meets continuously to discuss policy matters. Their attention, however, is focused on tax, borrowing, and monetary policies rather than on the details of government spending.

In the 1960s the attempt was made to plan government expenditures on a long-term basis. Emphasis was placed on the predetermination of expenditures, over a five-year period, in accordance with the expected (or desired) rate of growth of the economy. Planned expenditures are broken down by major functional categories such as defense, education, health and welfare, housing, and so forth, which are to serve as guides in the preparation of annual budgets. The functional breakdown is also intended to assist legislative discussion.

More recently, the elaborate systems of planning and control developed in the 1960s have been abandoned. There is less, and less detailed, forward planning of public expenditure. Forecasts are presented in cash terms, and the levels of inflation or cost increases contained within them are only partly disclosed. The primary disadvantages of the old system were that it allowed expenditures to be driven excessively by the burden of past commitments and to be blown off course by rises in the cost of public sector inputs. The new system has almost the opposite weaknesses: it induces short-term planning, in which capital expenditure is squeezed in favour of current spending, and allows the volume of outputs and inputs to be governed by optimistic assumptions about price increases. This development, intended to assist in public expenditure control by providing simple-to-understand criteria, has, according to some critics, reduced public understanding of the public expenditure process and therefore the accountability of government for its actions.

West Germany. West Germany differs somewhat from other countries in that there are unusual constraints on government borrowing and unusual reliance on countercyclical taxes and reserves. This is a direct result of Germany's history of extreme inflations. Detailed rules of budgetary behaviour apply to particular circumstances to determine whether deficits of borrowing are permitted.

The Federal Finance Plan

After a steep expansion in government expenditure in the 1950s and early '60s, new legislation was passed in 1967 to restructure the budgetary process. This provides for a five-year Federal Finance Plan that covers expenditures and receipts of the federal government, the *Länder* (states), and the local authorities for each year of the plan. The plan includes the budget for the present year, the draft budget for the next year, and estimates for each of the next three years. These financial plans are linked to a macroeconomic projection based on published target values for various economic indicators. The system ensures that expenditure and taxation are planned together for a five-year period and that countercyclical measures are also considered in such a medium-term framework rather than as panic responses.

Japan. Under Japan's 1947 constitution the Cabinet has the responsibility of preparing the national budget, which must then be submitted to the lower house of the Diet. Taxes can be imposed or modified only as prescribed by laws enacted by the Diet.

The budget is prepared on a fiscal-year basis by the budget division of the Ministry of Finance. The centre of the budget system is the general account, which theoretically includes all revenue and expenditure directly applicable to the overall fiscal operation of the government. There is also a system of special accounts for the operation of government enterprises and other special aspects of government finance. Theoretically, each special account

is self-balancing. In actual practice, however, there have at times been substantial deficits in the special accounts that have had to be covered by direct government appropriations, borrowings, and transfers of funds from one account to another.

Under the Public Finance Law of 1947, the general account of the national budget must be either balanced or in surplus. The government cannot increase its net long-term debt without special legislation, and then the increase must be tied to some specific investment use.

The Soviet Union. In the Soviet Union and some other countries having Communist governments, economic activity is either carried on by state enterprises or is subject to central control. The national budgets therefore have a much broader scope than in countries where most economic activity is in the private sphere. In the Soviet Union more than 90 percent of capital investment is financed by the government; in the United States the corresponding figure is less than 25 percent, and in the United Kingdom it is less than 50 percent. There are two main sources of revenue: the profits of state-owned enterprises and the turnover tax on sales of goods from these enterprises. The relative proportions of the funds drawn from enterprises under these two headings vary; there has been a long-term tendency for the share of the turnover tax to decline and for profit transfers to increase.

The budget under central planning

The budget is essentially a part of the national economic plan. It is drawn up by the Ministry of Finance and presented to the Supreme Soviet for consideration and approval every calendar year. Revenues and expenditures are usually in close balance. The budget is implemented by the Ministry of Finance, which scrutinizes the operations of the state enterprises in accordance with the economic plan.

BUDGETS OF OTHER LEVELS OF GOVERNMENT

Although the major budgetary decisions that affect the performance of the economy and the national debt are usually taken by the central government, most countries have local or state governments that are responsible for the provision of various services and have the authority to raise revenues through taxation or to borrow on their own account. This devolution of authority is greatest in the United States, where the majority of provision of civilian services is carried out at state or local levels and where states have a tradition of being individual decision-making units. In the United Kingdom, by contrast, local authority spending is constrained by rules set by the central government. Local authorities are also limited in their ability to borrow and to raise taxes, which are set by the central government. The budget of the European Economic Community is an example where authority for major spending, particularly for agricultural support, has devolved to a transnational body.

State and local budgets in the United States. Expenditures by state and local governments grew rapidly in the United States after World War II. Spending expanded particularly in the areas of education, health, and welfare services. Rising expenditures have been coupled with an expansion of the number of taxes used at this level and a widening of the traditional tax base from property. Although some local activity is financed by federal grants or by taxes that are shared between the federal and state level, the majority of expenditure at state and local levels is raised by locally determined taxes. Most federal grants are specific, and for particular and limited purposes. The majority of such grants concern education, income security (*e.g.*, public aid, housing assistance), and health programs such as Medicaid. The appropriate balance both of public provision and of revenue sources between federal, state, and local levels is the subject of continuing and vigorous debate.

Development of income tax by the states

The individual states pioneered the use of income taxation before World War I, but today the federal government is the major user of this form of revenue, with only limited scope for state income taxes. The decline of income taxes as a source of local finance at the state level was also hastened by problems with the definition of income and rising collection and compliance costs.

Diminishing taxation of income, coupled with congressional rejection of a general sales tax in the 1930s, led to a growth in the use of the retail sales tax; in the late 20th century most states use this as their major source of finance. Taxation of business income at the state level has also run into problems, not least because of its damaging effects on interstate trade; accordingly, although most states have tried such taxes, their importance remains limited. Property taxes, once the major source of state revenue, have, during the 20th century, become largely a local tax.

The 20th century has witnessed a large shift in expenditures on civil functions. In 1902 more than 70 percent of such functions were supported by local governments out of revenue raised predominantly from property taxes. But through the following decades, the proportion provided by local governments shrank to less than one-half, with the states taking on a larger share of the funding and the federal provision also increasing. During the period since World War II, the balance of state and local expenditure has shifted somewhat, with education increasing in importance (today it accounts for more than one-third of the state and local budget), while highways have increasingly become a federal function, through either direct federal funding or grants-in-aid to state and local governments.

Relations between the federal and state governments have been through a number of cycles in the postwar period. In 1948 grants to state and local governments were small, but they increased so dramatically during the 1960s that only the most dynamic local administrations could keep up. The allocation of resources came to depend more on the efficiency and wit of local administrators than on any particular need. A backlash to this situation took place during the 1970s, with a strong movement toward grants allocated by formula (loosely based on need) rather than application from the states. In 1972 Congress passed the State and Local Fiscal Assistance Act, which over a five-year period allocated some $30,000,000,000, one-third to state governments and two-thirds to local governments. This act, called general revenue sharing, continued into the 1980s although the amounts it allocated generally diminished after 1980. Only a fraction of the total amount of federal aid to state and local governments, estimated to exceed $100,000,000,000 per year by the mid-1980s, was provided by the act.

Federal grants to local government

Local government finance in the United Kingdom. The United Kingdom has a number of different levels of government, ranging from local parishes to the central government, and including districts (town, borough, or city), counties, and metropolitan counties. All of these levels have different traditional roles and are responsible for the provision of particular services. Historically, apart from direct contributions to finance the activities of parish councils and charges for local services, the only source of local revenue has been a property tax called rates.

Rates are levied, at levels determined by county and district authorities, on both households and businesses in proportion to the ratable value of the property they occupy. This ratable value, in principle revised periodically, is supposed to reflect the rental value of the property in a base year; increases in revenue are caused not by changes in ratable values but by steady increases in rate poundage (tax rate).

Use of ratable value

Reassessments of ratable values have proved politically unpopular and hence have been repeatedly postponed—the last revaluation in England and Wales occurred in 1973. Almost one-third of all households now receive rebates on grounds of low income, and rates paid for industrial and commercial premises considerably exceed those on domestic property. Many local authorities have perceived that only a small part of the overall tax burden falls on local electors and have not hesitated to implement substantial increases in the tax rate. The government has intervened to limit the rates that can be charged by high-spending authorities and has been actively pursuing alternative schemes of financing.

Rates today raise about half of all local authority expenditure, so that local authorities are very dependent on grants from the central government. These are determined according to complex formulas that relate them to needs

and resources in each local area. The central government now tries to limit local expenditure by imposing penalties on high-spending authorities through the grant system.

Budget of the European Economic Community (EEC). Each year, European countries that are part of the European Economic Community agree on a budget for the EEC. This covers expenditure on agricultural support, on regional and social development, and on the financing of various transnational agencies. The budget is presented in a series of drafts that are substantially revised following discussion among ministers from the various countries.

Finance for the budget of the EEC is provided by the common external tariff levied on all imports from outside the EEC, by levies raised on agricultural imports, and by a national contribution that is set, in most cases, as a fixed percentage of the "value-added base" (roughly the national income) of each member state. This is subject to a limit set at 1 percent by the Treaty of Rome, but which was subsequently raised to 1.4 percent.

The EEC budget is mostly devoted to financing the Common Agricultural Policy (CAP), which endeavours to protect farmers by guaranteeing them high prices for their output, often at as much as double the world price. This has led to overproduction of many goods and the accumulation of various surpluses, which are stored at community expense or sold at cheap prices, often to the Eastern bloc countries. The costs of this policy have been difficult to control, and the budget of the EEC, as a consequence, has lurched from crisis to crisis as the limits on agricultural expenditure have been steadily exceeded. (J.A.Ka./Ed.)

The economic functions of government

A significant development in the intellectual history of the 20th century has been the explicit recognition by economists, politicians, and the public at large of the importance of government in the operation of the economy. The national budget generally reflects the policy of the government toward the economy. Three main roles of the government in controlling the economy may be considered: the allocative function, the stabilization function, and the distributive function.

Government's main roles in controlling the economy

Over time, there have been considerable changes in emphasis on these different economic functions of the budget. In the 19th century, government finance was primarily concerned with the allocative function. The job of government was to raise revenue as cheaply and efficiently as possible to perform the limited tasks that it could do better than the private sector. As the 20th century began, the distribution function acquired increased significance. Social welfare benefits became important, and many countries introduced graduated tax systems. In the later interwar period, and more especially in the 1950s and '60s, stabilization was central, although equity was also a major concern in the design of tax systems. In the 1970s and '80s, however, the pendulum swung back. Once more, allocative issues came to the fore, and stabilization and distribution became less significant in government finance.

THE ALLOCATIVE FUNCTION

The allocative function in budgeting determines on what government revenue will be spent. Because a high proportion of national income is now devoted to public expenditure, allocation decisions become more significant in political and economic terms. At all times and in all countries the calls for expenditure on specific services or activities, or for more generous transfer payments, will always exceed the amount that can reasonably be raised in taxation or by borrowing. The debate about how these scarce resources should be allocated has continued for hundreds of years, and although numerous methods of deciding on priorities have emerged, it has never been satisfactorily resolved. In practice, most democracies contain a number of different factions that disagree on the proper allocation of resources and indeed the proper level of public sector involvement in the economy; the frequent change of national governments is related to the constant search for the right answers.

Public goods. Economists have sought to provide objective criteria for public expenditures through the so-called theory of public goods. It is generally recognized that some goods needed by the public cannot be provided through the private market. Lighthouses are a classic example. The costs of a lighthouse are such that no one shipowner will want to finance it; on the other hand, if a lighthouse is provided for one shipowner, it can be made available to all for no additional cost. Indeed it must be available to all, since there is no practical means of excluding ships from using the facility provided by the lighthouse, even if their owners have refused to pay for it. The only practical method of providing such services is by collective action.

Debate over paying for public works

If goods are to be provided in this way, rather than through the private market, it is immediately necessary to confront the twin problems of deciding how much to provide and who should pay for that provision. Even if all individuals wanted the service equally—as, perhaps, with lighthouses—their views on the extent of the service would be influenced by the allocation of the costs. Where different households may have different preferences, and some may not want the service at all—as, for example, with defense by nuclear weapons—these difficulties are compounded. Economists have tried to devise abstract voting schemes that would reconcile these difficulties, but these appear to have little practical application.

Moreover, others would challenge this whole approach to the problem. It would be absurd to say that the consumer has a taste for national defense and that it is the job of the government to satisfy it. The task of national leaders is to evolve a defense policy and persuade the public to accept it. Similarly, conservationists must attempt to awaken the public to the importance of parks and wildlife. In the context of public policy, the efficient allocation of resources consists not merely of distributing funds in the pursuit of given objectives but also involves determining the objectives themselves.

Genuine public goods pose severe problems for the national budget; it is very difficult to decide how far particular goods—the arts, national parks, even defense—should be supplied, and therefore no formal procedure of determination is likely to evolve. What should be given to each will continue to be the subject of intense political debate, with allocation changing as the government changes.

Merit goods. The concept of merit goods assists governments in deciding which public or other goods should be supplied. Merit goods are commodities that the public sector provides free or cheaply because the government wishes to encourage their consumption. Goods such as subsidized housing or social services, which predominantly help the poor, or health care services, which help the poor and elderly, are generally regarded as having considerable merit and therefore have a strong claim on government resources. Other examples include the provision of retraining schemes or urban regeneration programs.

Cost-benefit analysis. Once decisions have been made on how the limited national budget should be divided between different groups of activities, or even before this, public authorities need to decide which specific projects should be undertaken. One method that has been used is cost-benefit analysis. This attempts to do for government programs what the forces of the marketplace do for business programs: to measure, and compare in terms of money, the discounted streams of future benefits and future costs associated with a proposed project. If the ratio of benefits to costs is considered satisfactory, the project should be undertaken. "Satisfactory" means, among other things, that the project is superior to any available public or private alternative. Or, if funds are limited, public investment projects may be assigned priorities according to their cost-benefit ratios.

One difficulty with cost-benefit analysis is that every government agency has an incentive to estimate favourable ratios for its own projects. It must, after all, compete with other agencies for funds. No one can be certain as to the returns to be expected from an irrigation canal or a highway. Private investors have also been known to exaggerate their claims in appealing to stockholders, but they are generally subject to market sanctions that encourage them to err on the side of caution.

In addition to the possibility that cost-benefit analysis may be biased by the preformed views of those commissioning the study, there are other, more fundamental difficulties. Almost all proposals have effects that are difficult to value in monetary terms. The siting of a new airport brings problems of noise and property blight to local people and increases the risk that civilians may die in an accident. Putting a sensible value on human life has been a continuing difficulty for those carrying out cost-benefit analyses, even though every project does in fact affect probabilities of life and death. These problems are, of course, not confined to cost-benefit analysis. Additional expenditure on health service, or on road safety, or better housing, or heating old people's homes in winter all affect the number of people who die prematurely. The failure of cost-benefit analysis to provide answers to the problems of valuing life, or the quality of life, is a reflection of the wider problem confronting all decisions on public expenditure: the influence of subjective judgment.

The difficulty of valuing human life

Public ownership and privatization. Until the mid-1970s, the proportion of economic activity controlled by the government and the share of taxes in national income tended to increase in most countries. Since then, however, challenges to this growth in the role of government have become increasingly influential, and moves to privatization have been common.

The trend to privatization

There are several types of privatization. One involves the sale to private owners of state-owned assets, and this is most correctly called privatization. Publicly owned houses may be sold to their occupants. Commodity stockpiles may be reduced or disbanded. Increasingly, however, attention has been turned to the sale of publicly owned industries, thus reversing the move to nationalization that occurred, particularly in western Europe, around and after World War II.

Where the privatized industry operates in a competitive environment, no new problems arise. Singapore has privatized its airline system, for example, which now competes with a mixture of privately and publicly owned international airlines. Where privatization occurs but monopoly continues, however, there are new difficulties. Both Japan and the United Kingdom have privatized their telecommunications networks. Although in certain limited areas of telecommunications competition is possible—and has been allowed to develop in both the United States and Britain—technical and legal restrictions inhibit competition in many sectors of the industry.

Regulation is necessary, therefore, to restrict the freedom of privatized monopolies, or near monopolies, to raise prices and to exploit consumers in other ways. In the United States, which has by far the longest history of regulating private utilities, such regulation has normally limited the rate of return that they earn to what is considered a fair level. A disadvantage of this is that it may give the industry no greater incentive to increased efficiency than would exist in public ownership, since higher costs can be passed directly onto consumers. There have been experiments, therefore, with other forms of regulation, which seek to strike a balance between incentives for better performance and the ability to exploit consumers.

A further problem for such regulation is that utilities and similar industries normally operate in both competitive and monopoly markets. They may be inclined to use their monopoly power in some areas to gain unfair competitive advantages in others. Despite these difficulties, an increasingly wide range of industries, ranging from water supply to airports, are now considered candidates for privatization.

Privatization can also mean the dismantling of existing statutory restrictions on competition. State activities are often protected by legal prohibitions on competing private enterprise. West German railways, for example, are entirely state-owned, and the law not only prevents competing railroads but severely restricts coach services and limits competitive trucking. The dismantling of such restrictions is seen as one method of improving the efficiency of state concerns.

Another demand of privatization is the contracting out of publicly provided services. U.S. municipalities have often entrusted activities such as refuse collection, and in some cases even fire service, to private contractors, and European countries are increasingly experimenting with similar schemes. These possibilities demonstrate that a service may be government-financed but not necessarily provided by the government; if extended more widely, the concept could yield a different view of the economic role of the state.

While the objective of privatization is often to increase the efficiency of government activities, its implementation may also have important effects on government revenue. Any savings that result from lower costs lead directly to lower tax rates. Where budgeting procedures do not distinguish between capital and current transactions, the proceeds of privatization sales provide a once-and-for-all boost to revenues. The availability of this source of funding for state activity has given an artificial attractiveness to privatization, especially in the United Kingdom. If an industry is sold for the present value of its expected earnings, and these earnings are the same in public and private ownership, privatization should have no net impact on public finances. If it is expected to be more efficient in the private sector, government finance, on balance, gains. If it is sold for less than the maximum revenue that would be obtained—and this is often the case, either because of the difficulty of selling assets as large as nationalized industries, or because the government wishes to secure a wide dispersion of share ownership—the impact is likely to be negative.

Privatization and government revenue

Other forms of government intervention. Government spending is not the only way in which government allocates resources. Its regional policies will determine whether domestic and overseas investors build factories in particular places, while its taxation policies will determine whether they build them at all. Government competition and merger policies affect the structure of industry and commerce, while regulatory activities—setting the number of hours shops may be open or who may buy cigarettes—have profound effects on commercial activities.

Government also affects allocations by setting the legal and administrative framework within which the economy functions. It may specify minimum wage levels or control the siting of new ventures and the activities of existing ones. Such activities of government profoundly affect the allocation of resources, but they are rarely monitored or subject to serious control.

THE STABILIZATION FUNCTION

Stabilization of the economy (*e.g.,* full employment, control of inflation, and an equitable balance of payments) is one of the goals that governments attempt to achieve through manipulation of fiscal and monetary policies. Fiscal policy relates to taxes and expenditures, monetary policy to financial markets and the supply of credit, money, and other financial assets.

History of stabilization policy. The use of fiscal and monetary policy as a means of stabilizing the economy is relatively recent, for the most part a development of the period after World War II. During the 19th century the only stabilization policy was that associated with the international gold standard. Under the gold standard, if a deficit occurred in a country's balance of payments, gold tended to flow out of the country. To counteract this process, the monetary authorities would raise interest rates and stiffen credit requirements, causing a fall in prices, income, and employment; this in turn led to a reduction in imports and an expansion of exports, thus improving the balance of payments. If a country had a surplus in its balance of payments, gold tended to flow in; this meant that the interest rate fell and the supply of money and credit was increased. As a consequence, imports were stimulated and exports discouraged so that the surplus in the balance of payments tended to disappear. The adjustment mechanism also included another important element: capital movements between countries. When interest rates fell in surplus countries and rose in deficit countries, mobile international financial capital tended to flow from the former to the latter, contributing to the elimination of deficits and surpluses in the balance

The gold standard as a tool of stabilization

of payments. The working of this mechanism was partly automatic and partly the result of deliberate actions by the monetary authorities in each country.

In this form of stabilization policy, external stability was achieved at the cost of stability in the domestic economy: fluctuations in domestic prices, incomes, and employment functioned as the levers for bringing about equilibrium in the balance of payments. Occasionally governments attempted to reduce the impact of this mechanism on the domestic economy, particularly on the price level. In particular, governments in some surplus countries took "sterilization actions" to prevent the gold inflow from increasing the supply of money and credit to the maximum extent. This could be done if the central bank offset its purchases of foreign exchange and gold with sales of government securities on the domestic credit market.

Stabilization policy of the 1920s and '30s

A somewhat more ambitious type of stabilization policy emerged in the period after World War I. During the late 1920s and early 1930s the need to reduce unemployment acquired more urgency. Previously, the exchange rate, the balance of payments, and occasionally the price level had been considered more important than the situation in the labour market. During the 1920s unemployment in Great Britain rose to very high levels (between 20 and 30 percent of the labour force). Consequently, there was much discussion of whether employment could be increased by actions of the public authorities. At first, the discussion in Great Britain centred on the feasibility of public works programs as a means of putting men to work; there was a growing belief that these programs might also be a good means of raising the general level of economic activity through their effect on purchasing power. Some maintained that budget deficits would also raise the level of economic activity. An active part in this discussion was taken by the economist J.M. Keynes, and also by the Liberal Party, which in 1928 published proposals for government intervention entitled *Britain's Industrial Future.*

The first countries to adopt the new policies were Sweden and Germany. When the Nazi Party took power in Germany in 1933, its rearmament policies helped to reduce unemployment and to stimulate the economy. In Sweden, the new Social Democratic government attempted in more modest ways to expand the economy and ease unemployment through increased government expenditures in 1932–33. In the United States, a very limited attempt was made by the administration of Pres. Herbert Hoover; but Franklin D. Roosevelt made a more aggressive effort with such projects as the Works Progress Administration (WPA), which carried on its payroll an average of more than 2,000,000 workers per year from 1935 to 1941. Unemployment, however, persisted at a high level until World War II, although there was a significant drop from a level of about 25 percent in 1933.

Stabilization theory. The new stabilization policy needed a theoretical rationale if it was ever to win general acceptance from the leaders of public opinion. The main credit for providing this belongs to Keynes. In his *General Theory of Employment, Interest and Money* (1935–36) he endeavoured to show that a capitalist economy with its decentralized market system does not automatically generate full employment and stable prices and that governments should pursue deliberate stabilization policies. There has been much controversy among economists over the substance and meaning of Keynes's theoretical contribution. Essentially, he argued that high levels of unemployment might persist indefinitely unless governments took monetary and fiscal action. At that time he believed that fiscal action was likely to be more effective than monetary measures. In the deep depression of the 1930s, interest rates had ceased to exert much influence on the ways in which owners of wealth disposed of their funds; they might choose to hold larger cash balances instead of spending more money as the traditional theory had suggested. Nor were investors inclined to take advantage of low interest rates if they could not find profitable uses for borrowed funds, particularly if their firms were already suffering from excess capacity. Keynes's pessimistic view of monetary policy had a strong influence on economists and governments during and immediately after World War II, with the result that monetary policy was not tried very much during the 1940s. It was often forgotten during the policy discussions of the time that Keynes's views on the efficacy of monetary policy were related to the particular situation of the 1930s.

The contribution of J.M. Keynes

Another influential idea embodied in Keynes's writing was that of economic stagnation. He suggested that in the advanced industrial countries people tended to save more as their incomes grew larger and that private consumption tended to be a smaller and smaller part of the national income. This implied that investment would have to take a continually larger share of the national income in order to maintain full employment. Since he doubted that investment would rise sufficiently to do this, Keynes was rather pessimistic about the possibility of achieving full employment in the long run. He thus suggested that there might be some permanent tendency to high levels of unemployment. This also had considerable influence on economic policy during the early postwar period; it was some time before those in decision-making positions realized that inflation, rather than stagnation and unemployment, was to be the main problem confronting them.

The desirability of pursuing policies to maintain high levels of employment was generally accepted in most industrial countries after the war. In 1944 the British government stated in its White Paper on Employment Policy that "the government accept as one of their primary aims and responsibilities the maintenance of a high and stable level of employment after the war." One of the most influential British economists at this time was Sir William Beveridge, whose book *Full Employment in a Free Society* had a strong impact on general thinking. Similar ideas were expressed in the United States in the Employment Act of 1946, which stated: "The Congress hereby declares that it is the continuing policy and responsibility of the Federal Government to . . . promote maximum employment, production and purchasing power." The Employment Act was less specific as to policy than the British government's White Paper, but it established a council of economic advisers to assist the president and called upon him to present to every regular session of Congress a report on the state of the economy. The president was also required to present a program showing "ways and means of promoting a high level of employment and production." Similar programs were adopted in other countries. In Sweden in 1944 the Social Democrats published a document somewhat similar to the British White Paper, and other such declarations were made in Canada and Australia.

The acceptance of full employment as a government goal

Fiscal policy. Fiscal policy attempts to control the actions of individuals and companies by means of spending and taxation decisions. On the expenditure side, it can achieve this by spending money in ways—for example, on construction projects—that stimulate other activity, while on the taxation side it can affect work, investment, or production decisions by changing tax rates and levels. Fiscal policy thus has two major components: an overall effect generated by the balance between the resources the government puts into the economy through expenditures and the resources it takes out through taxation, charges, or borrowing; and a microeconomic effect generated by the specific policies it adopts. Both are important in stabilizing the economy.

The two major components of fiscal policy

Overall fiscal policy involves the government in deciding whether it should spend more than it receives or less. The development of countercyclical fiscal policies in the post-World War II period reflected the explicit attempt by some governments to protect their population from world recessions by deliberately spending additional money at appropriate times. Experience with countercyclical fiscal policy has been disappointing; in many cases, the lag between identifying the problem and fiscal response has been too long, with the result that a fiscal boost coincided with the next boom, while a contraction might coincide with the next recession. Fiscal policies that were intended to be countercyclical could end up exacerbating the original problems.

Another facet to fiscal policy is a government's attempt to guide the development of the economy by more specifically targeted policies. Thus most countries have from

time to time attempted to cushion particular areas from the effects of a decline in their dominant industry by regional policies, to affect labour supply and demand by taxation, and to change the pattern of consumer purchases by changes to indirect taxes. These policies sometimes backfire as unforeseen consequences and interactions occur.

The simple notion of budget balance, although widespread, can be seriously misleading to one who attempts to decide whether a government is being expansionary or contractionary at a particular time. If nothing else were happening, and there were no inflation, no changes in unemployment or exchange rates, and the country were to have a constant population of all ages, then the government's fiscal position, or stance, might be said to be neutral (neither expansionary nor contractionary) if spending were an exact match of taxation, charges, and profits on public sector activities. (Even in such a case, however, if it were pursuing specific microeconomic policies, its neutrality might hide significant effects on the behaviour of the economy.) This notion has led many countries to believe that fiscal position is appropriately measured by the size of public borrowing, because this measures the difference between the amount government spends and the amount it receives.

Neutral fiscal position

The recognition that simple budget balance (not accounting for inflation) may not in fact be neutral when other things are changing has led to a number of suggestions for more sophisticated measures of fiscal position. The full-employment budget surplus suggested by the Council of Economic Advisers in the United States, for instance, attempts to adjust the simple measure of budget deficit or surplus in reaction to the effects of deviations from a level of unemployment that it regards as "normal" or "full." The argument for this kind of adjustment is that high levels of unemployment cause increased benefit payments and reduced tax receipts that are abnormal, and if the government were to try to maintain a simple budget balance at times of high employment, this would require a large contraction in the other activities it supports. A simple deficit, then, may be a surplus on a full-employment basis, and government action may be severely contractionary despite positive levels of borrowing.

Another type of suggested adjustment recognizes that inflation erodes the real value of public debt. The neutral simple budget balance, it is argued, only requires that the government maintain its real asset position. If inflation is eroding the real value of existing debt, then the government may borrow to an adjusted, or revised, level before its actions actually reduce public assets.

Although it has come to be recognized that a simple budget deficit or surplus does not adequately reflect a government's fiscal position, no country directly employs measures revised for unemployment and inflation in deciding on countercyclical policies. This is partly because they are more difficult for politicians to understand and partly because it is genuinely difficult to decide on the precise form they should take. Whatever the reason, many countries, even at times of high inflation and unemployment, continue to focus on the simple budget balance measures. The United Kingdom, for example, continued in 1980–81 to attempt to reduce public borrowing during a serious world recession and ran an adjusted surplus. This procyclical policy is blamed by many as a cause of the high levels of unemployment that subsequently prevailed in that country.

The heyday of fiscal stabilization policies was, however, the 1950s and '60s. In the 1970s governments became increasingly concerned about inflationary pressures, and important disturbances, particularly the oil crisis, disrupted world economies. Stabilization became a less important policy goal and one that governments were increasingly unable to achieve. Monetarist economic theories acquired increased influence. The primary economic issues determining fiscal policies once again became the more traditional concerns of allocation and distribution.

Monetary policy. Although the governmental budget is primarily concerned with fiscal policy (defining what resources it will raise and what it will spend), the government also has a number of tools that it can use to affect the economy through monetary control. By managing its portfolio of debt, it can affect interest rates, and by deciding on the amount of new money injected into the economy, it can affect the amount of cash in circulation and, therefore, indirectly affect prices and other economic variables. In recent years, governments, discouraged by past failures with fiscal manipulation, have turned to monetarist policies to attempt control of the economy.

Control of interest rates through debt management

At its simplest, monetarist theory postulates that in the economy there is a fixed amount of money, which circulates at a given velocity. This money is then available to finance the various transactions carried out in the economy at the prevailing prices. Under these circumstances, according to the theory, control of the price level can be maintained by controlling the amount of available money.

Although a desire to control inflation has been at the heart of the recent rise to prominence of monetary policies in many countries, monetary policy can be used to affect a number of different facets of economic behaviour. In time of unemployment the central bank may stimulate private investment expenditure, and possibly also household spending on consumer goods, by reducing interest rates and taking measures to increase the supply of credit, liquid assets, and money. The customary tools for doing this are open market operations, the discount rate of the central bank, and cash reserve requirements for commercial banks.

In open market operations the central bank buys government securities—bonds and treasury bills—from the private sector. The effect is to reduce interest rates by bidding up bond prices. The sellers of the government securities obtain cash that they deposit in the banks, thus increasing the cash reserves of the banks and enabling them to expand credit to private borrowers; this in turn causes interest rates in the private sector to fall and the terms of credit to become easier. In response, firms are likely to increase their investment expenditures, and households are likely to spend more on consumer goods.

The second tool of monetary policy, the discount rate of the central banks, is often used together with open market operations. This is the interest rate at which commercial banks can borrow funds from the central bank. If the discount rate is reduced, banks become more willing to extend credit to private borrowers because they can obtain funds themselves on easier terms. In many countries, changes in the discount rate tend to be followed by similar changes in the interest rates charged by banks to their borrowers.

The third tool of monetary policy, that of the cash reserve requirements (and, in some countries, certain types of government securities) for commercial banks, provides that banks must maintain money balances (in the form of deposits in the central bank) at a certain proportion of their liabilities. This means that the banks cannot expand their earning assets such as government securities and private loans, beyond that point. If the government reduces the reserve requirements, the banks can expand their loans further, thus increasing the total volume of credit outstanding.

Monetary policy, like fiscal policy, may also be used to combat inflationary tendencies by reversing the above measures; the central bank will then sell government securities (thereby increasing interest rates and reducing the supply of private credit and money), raise the discount rate, or increase reserve requirements.

Stabilization policy problems. A broad distinction may be made between two types of stabilization policies: discretionary and automatic. Discretionary policies involve deliberate actions taken by the authorities, such as open market operations, changes in discount rates and reserve requirements, and changes in tax rates or government expenditures. Automatic policies put reliance on built-in stabilizers that function without any deliberate intervention by the authorities. In the monetary field, for example, an increase in commodity prices tends to reduce the real value of financial assets, and if the government does nothing to offset this by increasing the volume of financial assets in the system, private spending will tend to decline. On the fiscal side, the main automatic stabilizer is the

Discretionary and automatic policies

relation between tax revenues and cyclical changes in the economy. During booms, tax revenues rise and the need for expenditures on unemployment compensation decreases, channeling a larger proportion of the national income into government coffers; these effects are accentuated if the tax system is progressive because tax revenues rise more rapidly than money incomes. Provided that the government does not raise its expenditures along with the increased revenues, the budget tends to have a braking effect on private expenditure in boom times and an expansionary effect in times of recession.

The problem of time lags. There has been much discussion over the merits of discretionary policies as against automatic stabilizers. One advantage of automatic stabilizers is that the effects occur without the necessity of government action, which means that there is no delay, or lag, because of political controversies, administrative problems, or difficulties in determining whether the time has come to act. There are three types of lag in economic policy: the recognition lag, the decision lag, and the effect lag.

The three types of lag

The recognition lag is the time it takes for the authorities to discover the need to make a change in economic policy. The reasons for this type of lag are that statistical information is often somewhat behind the event and that it is sometimes difficult to distinguish between random fluctuations and fundamental shifts in economic trends. Governments prefer to wait until there is certainty that, say, an increase in unemployment is not a passing thing.

The decision lag is the period between the time when the need for action is recognized and the time when action is taken. Although the recognition lag is presumably of about the same duration for both monetary and fiscal policies, the decision lag is usually considerably shorter for monetary policy than for fiscal policy. The central bank can change monetary policy almost overnight, whereas a change in fiscal policy is more complex, both politically and administratively. In many countries changes in income taxes, for example, can be made only at the beginning of a calendar year; such changes are often complicated by political discussions in the legislative body.

The effect lag is the amount of time between the time action is taken and an effect is realized. Monetary policy involves longer delays than fiscal policy; the time between a change in monetary policy and its ultimate effect on private investment may be between one and two years.

Some economists argue that the sum of all the lags is so long and uncertain that the best strategy is not to take any action; by the time the effects occur the economic situation may be radically different. Some countries have tried to shorten the lags in fiscal and monetary policy. One way to reduce the recognition lag is to improve the forecasting techniques, for example, by using sophisticated questionnaires or computerized econometric models.

In order to reduce the decision lag in fiscal policy, some countries have given the authorities power to take limited action without the prior consent of the legislature. In the United Kingdom the government introduced a regulation that allowed it to make immediate changes in tax policy. In Belgium and West Germany the governments also have some discretionary powers to change tax rates without first asking the legislature. In most countries, however, the legislative bodies have been reluctant to give up control of the budget, and increasing skepticism about the effectiveness of stabilization policy has led to a retreat from frequent small adjustments to fiscal policy.

Attempts to shorten the effect lag of fiscal policy have produced new policy tools. Some countries now use systems of taxes or subsidies to influence business investment within a relatively short time. Attempts have also been made to reduce the effect lag in monetary policy. Some countries have tried using various tools of credit rationing rather than relying on traditional measures such as open market operations. But the effect lag is still a serious problem for monetary policy.

The multiplicity of policy goals

Conflicts among goals. Perhaps the most serious unsolved problem of stabilization policy is the multiplicity of goals that policymakers must consider. Every government has aims other than stabilizing the economy. First, it must stay in power—a need that is likely to limit the alternatives open to stabilization policy, particularly in periods of prosperity immediately before elections. Second, some monetary and fiscal actions impinge on particular groups in society, and governments may wish to avoid what appear to be discriminatory policies. Third, a policy designed to achieve one element of stabilization, such as full employment, may prevent the achievement of another.

The conflict between full employment and price stability seems to arise in two different sets of circumstances. Often wage increases that are made in the normal collective bargaining process are greater than the increases in labour productivity (or output per man-hour); such wage increases tend to increase the cost of production and to force prices upward. The government is then confronted with a choice between two unpleasant alternatives. One is to allow the general price level to rise approximately in proportion to the increase in production costs; the other alternative is to try to hold prices down by taking measures to restrict aggregate demand, thus making it difficult for firms to shift their increased costs to the consumer through higher prices. The latter alternative means increased unemployment. Many governments have been confronted with exactly this choice of alternatives. Wage gains made in collective bargaining have forced them to choose between allowing prices to move upward or attempting to hold prices stable at the cost of greater unemployment.

Another reason for the conflict between full employment and price stability is the tendency of wage increases to accelerate when the level of employment rises and the number of job vacancies increases. In other words, as the economy approaches full employment wages tend to rise at an increasing speed. As prices begin to rise, the conflict between full employment and price stability may be further exacerbated by the expectation that they will rise still further; this may, for example, induce employees and their organizations to press for greater wage increases than they otherwise would in order to compensate for the expected price increases.

Stabilization policy and the balance of payments

Another conflict in policy may arise with respect to the balance of payments. When the economy is in a period of boom, there is a tendency for imports to increase, and sometimes for exports to decrease as well, with obvious difficulties for the balance of payments. The crisis may be heightened by short-term capital movements if buyers and sellers of foreign exchange expect that there may be a devaluation of the country's currency. This has caused much difficulty for many countries in the period since World War II. In Britain and Denmark, notably, periods of boom have usually been accompanied by balance-of-payments problems. When that occurs, the government must sooner or later take restrictive actions that slow the economy down and increase unemployment; if speculation in the currency is already under way, it may be necessary to pursue the restrictive policy far into the next recession. The problem is accentuated if there have been substantial price increases during the boom that have reduced the country's ability to compete with other countries. It is ironic that a temporary improvement in the employment situation may, if it leads to an accelerated increase in the price level, serve to create greater unemployment in the future, when restrictive actions become necessary for balance-of-payments reasons.

Ways of reducing conflicts of policy

Attempts have been made to eliminate these conflicts of policy. One remedy is "incomes policy," direct efforts by the government to prevent employers and unions from raising prices and wages. Various methods have been tried. The most moderate is the so-called guideposts system, under which the government announces the need for restraints on wage increases and perhaps also sets targets to guide unions and management; this was attempted in the United States in the early 1960s. In Sweden, responsibility for limiting wage increases has been assigned to labour-management organizations where bargaining takes place in a centralized fashion. A more interventionist approach is for the government to enter the bargaining process and try to persuade unions to limit their wage demands. The government may go still further and announce a wage freeze, or even a system of wage and price control. In The

Netherlands, the courts have occasionally been empowered to set wages, but the resulting decisions have often been uncoordinated with the rest of stabilization policy.

Incomes policies have sometimes succeeded for short periods. Generally, however, public refusal to accept the restraints has eventually led to their collapse. In the United States, the guideposts broke down during the boom of the mid-1960s, and attempts at incomes policy in Sweden and Britain have not been notable for their success. Even in The Netherlands later attempts to impose the system have failed to limit the rate of wage increase.

The question of governmental competence. Governments have displayed serious deficiencies in their ability to handle stabilization policy. Political leaders often lack economic information and understanding, and their economic advisers find it difficult to explain the economic situation to them and to apprise them of the relevant tools. There are also a variety of political inhibitions against taking action. One consequence is that what is designed to be a countercyclical policy becomes a procyclical one; instead of stabilizing the economy it tends to destabilize it. The postwar experience in Britain is held by some to demonstrate the deficiencies of government in handling monetary and fiscal policy. In time of boom the government often followed an expansionary course; when a balance-of-payments crisis developed it then took restrictive action—too late—and pushed the economy into deeper recession than would otherwise have occurred. On the basis of this experience, some economists have argued that a policy that did not attempt to counter the short-run swings in the economy would have been more successful in achieving stabilization. They maintain that the authorities should concentrate on letting the volume of money and credit increase steadily at a rate dictated by the long-term growth trend of the economy. Those who hold this view believe that capitalist economies are inherently stable, that crises are usually the result of bad policies on the part of the public authorities. Most economists do not share their optimism as to the stability of the economy if left alone; they continue to believe that governments must seek better tools for the purpose of short-run stabilization.

Avoidance of a post-World War II depression

Experience in selected countries. The application of full-employment policies after World War II was made more difficult by the fact that the postwar situation was radically different from that of the 1930s, when much of the policy thinking had been done. Most governments and their advisers expected a depression after the war, but it never materialized. One explanation is that the reallocation of resources from military to civilian uses proceeded more smoothly than expected. Another explanation is that the consumers spent a larger part of their disposable income than they had been observed to do in the 1930s, upsetting some of the statistical projections based on empirical data from those years. A third explanation, which applies perhaps to the years after 1948, was the Cold War between the United States and the Soviet Union, which raised defense spending in many countries.

The period of the late 1940s and early 1950s proved to be characterized by tendencies to inflation rather than to unemployment. Governments were slow to realize this and to shift their emphasis from employment-creating policies to anti-inflationary policies. The fact that governments had accepted, to a large extent, the belief that monetary policy was not very important made it difficult for them to combat the tendencies to inflation. In most countries a passive, even expansionary, monetary policy was in effect; interest rates were kept down and the supply of money was allowed to grow faster than would have been consistent with stable prices. Inflationary tendencies were further stimulated by the Korean War and the great increases in raw material prices that accompanied it.

Developments in the 1950s

During the 1950s several important developments influenced the attitudes of governments toward stabilization policy. Most of the economic controls engendered by the war were removed, particularly in international trade and finance. The western European countries were in a period of rapid economic growth. With the removal of direct controls on prices, imports, and building investment, governments began to develop and refine the tools of monetary and fiscal policy. In most countries the passive attitude toward monetary policy disappeared during the early 1950s; there was increased interest in more flexible monetary management. Interest also grew in developing a systematic fiscal policy that would offset the cyclical swings in production and employment. The most energetic attempts to devise a countercyclical fiscal policy were made in Britain and Sweden. Other European countries and the United States placed more reliance on monetary policy.

In the United States, a contributing factor in the revival of monetary policy was a theoretical reformulation that took place among monetary and banking experts. This was the so-called availability theory of credit; it held that monetary policy had its effect on spending not only directly through interest rates but also by restricting the general availability of credit and liquid funds. It was argued that even rather small changes in the rate of interest for government securities could have a considerable impact on the supply of private credit; if the supply diminished, this would induce banks and other financial institutions to stiffen their credit standards and ration credit to their customers; this in turn, it was argued, would tend to curb investment and thus have a braking effect on the economy. Similar ideas were at work in other countries, but with more emphasis on limiting the availability of credit through credit rationing, loan ceilings, control of private bond issues, regulation of installment credit for the purchase of durable consumer goods, and so on.

Serious attempts have been made to put a countercyclical monetary policy into practice in most advanced industrialized countries since the middle of the 1950s. In some, such as the United States, the emphasis has been, as suggested above, on changes in the interest rate and in the supply of money and credit; in others, such as France, Italy, and Japan, the emphasis has been on the rationing of credit by the central bank.

Fiscal policy has found less use than monetary policy in efforts to control cyclical fluctuations in the economy. It has been most favoured in Britain, the Scandinavian countries, and The Netherlands. There are specific situations in which fiscal measures have been used to stimulate the economy in other countries, as in Belgium and West Germany during the recessions of 1958 and 1962. Another example is the postponement of certain military expenditures in the United States as an anti-inflationary measure during the boom of the mid-1950s, and, most notably, the tax cuts passed by Congress in 1964 and 1981 as a stimulus to economic expansion. Several countries have taken restrictive fiscal actions to overcome balance-of-payments crises, including France and Finland on various occasions. During the later 20th century there was an increasing tendency to employ fiscal policies in the short run, partly in order to assist monetary policy in solving cyclical problems.

THE DISTRIBUTIVE FUNCTION

Virtually everything that a government does has some effect on the distribution of income or wealth at the various levels of society. Improvements in health care facilities benefit the sick, the old, and those about to have children. An increase in taxes on tobacco and beer affects the poor disproportionately, while an increase in capital taxes similarly affects the rich. Even regulatory and legislative activity benefits one group out of proportion to another. The redistributive consequences of the governmental budget can be reflected in a variety of ways; sometimes they are explicit and sometimes they are cited in the debate that follows the presentation of a budget. Usually, however, these consequences are hidden, unintended, and imperfectly understood.

Incidence of taxation and expenditure. The incidence of taxes is a subject that has generated much academic debate. It is usual to distinguish between the legal incidence of a tax and its effective, or final, incidence. The legal incidence is on the person or company who is legally obliged to pay the tax. Effective, or final, incidence refers to who actually ends up paying the tax; if, for example, the whole of a sales tax can be passed on in higher prices to the consumer, then consumers bear the final incidence of the tax.

The legal and final incidence of tax

Whether the final incidence of the tax is on those who actually pay it depends on their market power relative to the people with whom they trade. A payroll tax, for example, is likely to be reflected in lower wages if labour is mobile and in plentiful supply, but may be borne, at least in part, by the employer if there are labour shortages. Similarly, if a manufacturer is facing intense international competition, his ability to pass on any increase in a sales tax is limited; on the other hand, his ability to do so is much greater if he is the sole supplier of a good.

Arguments for income redistribution. Although governments do affect the distribution of resources in numerous ways, this is often a by-product of the other things they are trying to do. It has been long debated whether or not governments should seek explicitly to redistribute income from the rich to the poor and, if so, to what extent. More generosity to the poor, whether through higher benefits or through a more progressive tax system, means a higher tax burden on richer people, with, it is argued, consequent effects on work effort and on other behaviour. The appropriate degree of redistribution has been the subject of an extensive literature on optimal taxes, but economists generally agree that the final determination must be through the political process.

Redistribution as a means of eliminating hardship

Economists point to a number of arguments favouring explicit as opposed to indirect methods of redistribution. The primary argument is that these would provide a more efficient means of eradicating severe hardship, suffering, or starvation. Left to itself, the market economy creates casualties among those who lack the skills to participate fully or those who have failed to generate sufficient resources to tide them over into old age. Countries have evolved programs for the prevention of severe need, although the definition of an acceptable minimum standard of living is typically more generous in European countries than in, for example, the United States, and this is reflected in the higher share of public expenditure in those countries. In most countries the definition of poverty, as measured by the level to which the state benefit system brings everyone's income, has moved from being "absolute" (determined by the minimum requirement of food, clothing, and shelter) to a more relative concept, which allows the poor to share in real rises in living standards.

The second argument for redistribution is that overall social welfare is thereby increased. An additional dollar makes more difference to the standard of living of someone earning $100 per week than to that of someone earning $1,000. Even if everyone has income above an agreed minimum level, there is a case for redistribution from the rich to the not-so-rich. The extent to which this should be pursued depends partly on the perceived distortions the redistribution would cause and partly on how much more value the far more numerous not-so-rich can squeeze out of each additional dollar.

Other arguments for redistribution occur where the market fails to allow individuals to redistribute between periods in their own lives. The classic example is that people tend to have their periods of highest expenditure (while bringing up children) at the points of minimum income (early in life). Those families with little or no access to credit markets can do very little about this, which has been used as one argument for redistribution toward those bringing up children. A second argument contends that children convey benefits to society as a whole, so that parents should be rewarded for creating a public good. This argument would, of course, have little strength in countries with serious problems of overpopulation.

A final group of arguments also concerns market failure. If particular areas or occupations have declined and the work force has not adjusted to this decline by moving to other areas or through retraining, then some subsidy to cushion the recessive effects might be considered appropriate. Most countries redistribute from better-off regions to those that have declined, or they allocate funds for specific programs designed to help particular groups.

(A.Li./J.A.Ka.)

BIBLIOGRAPHY. Studies of the governmental budget include JESSE BURKHEAD, *Government Budgeting* (1956); E.S. KIRSCHEN *et al., Economic Policy in Our Time,* 3 vol. (1964); J.R. HICKS, *The Social Framework: An Introduction to Economics,* 4th ed. (1971); UNITED STATES. CONGRESS, *Planning, Programming, Budgeting: Inquiry* (1970); DAVID J. OTT and ATTIAT F. OTT, *Federal Budget Policy,* 3rd ed. (1977); PETER C. SARANT, *Zero-Based Budgeting in the Public Sector: A Pragmatic Approach* (1978); and WILFRED BECKERMAN, *An Introduction to National Income Analysis,* 3rd ed. (1980). *Survey of Current Business* (monthy), published by the Bureau of Economic Analysis of the United States Department of Commerce, together with its supplement, *The National Income and Product Accounts of the United States: Statistical Tables* (irregular), provides practical analyses. The Brookings Institution annual *Setting National Priorities* explores U.S. budgetary developments; and the London Institute for Fiscal Studies does the same for the United Kingdom in its *IFS Report Series.*

Components of the budget and their interrelation are explored in RICHARD A. MUSGRAVE, *Fiscal Systems* (1969, reprinted 1981); HUGH HECLO and AARON WILDAVSKY, *The Private Government of Public Money,* 2nd ed. (1981); ALAN T. PEACOCK and JACK WISEMAN, *The Growth of Public Expenditure in the United Kingdom,* 2nd rev. ed. (1967); LEO PLIATZKY, *Getting and Spending: Public Expenditure, Employment, and Inflation,* rev. ed. (1984); CARL S. SHOUP, *Public Finance* (1969); HAROLD F. WILLIAMSON (ed.), *The Growth of the American Economy,* 2nd ed. (1951); ALAN S. BLINDER *et al., The Economics of Public Finance* (1974); CHARLES L. SCHULTZE, *The Politics and Economics of Public Spending* (1968); J.A. KAY and M.A. KING, *The British Tax System,* 3rd ed. (1983); CAROLYN WEBBER and AARON WILDAVSKY, *A History of Taxation and Expenditure in the Western World* (1986); MERVYN A. KING and DON FULLERTON (eds.), *The Taxation of Income from Capital: A Comparative Study in the United States, the United Kingdom, Sweden, and West Germany* (1984); JOHN F. DUE and JOHN L. MIKESELL, *Sales Taxation: State and Local Structure and Administration* (1983); CHARLES E. WALKER and MARK A. BLOOMFIELD (eds.), *New Directions in Federal Tax Policy for the 1980s* (1983); MARK ASHWORTH, JOHN HILLS, and NICK MORRIS, *Public Finances in Perspective* (1984); HENRY C. ADAMS, *Public Debts: An Essay in the Science of Finance* (1887, reprinted 1975); JOHN MAYNARD KEYNES, *How to Pay for the War* (1940); TILFORD C. GAINES, *Techniques of Treasury Debt Management* (1962); WARREN L. SMITH, *Debt Management in the United States* (1960); EDWARD NEVIN, *The Problem of the National Debt* (1954); HENRY C. MURPHY, *The National Debt in War and Transition* (1950); A. JAMES HEINS, *Constitutional Restrictions Against State Debt* (1963); JAMES M. BUCHANAN, *Public Principles of Public Debt* (1958); and JAMES M. FERGUSON (ed.), *Public Debt and Future Generations* (1964, reprinted 1982).

The budgetary process at both national and local levels is discussed in RICHARD A. MUSGRAVE and PEGGY B. MUSGRAVE, *Public Finance in Theory and Practice,* 4th ed. (1984); ARTHUR SMITHIES, *The Budgetary Process in the United States* (1955); AARON WILDAVSKY, *The Politics of the Budgetary Process,* 4th ed. (1984); SIR HERBERT BRITTAIN, *The British Budgetary System* (1959); SAMUEL BRITTAN, *Steering the Economy: The British Experiment* (1971); R.W. DAVIES, *The Development of the Soviet Budgetary System* (1958, reprinted 1979); DAVID NOVICK (ed.), *Program Budgeting: Program Analysis and the Federal Budget,* 2nd ed. (1969); RICHARD A. MUSGRAVE (ed.), *Essays in Fiscal Federalism* (1965, reprinted 1977); WALLACE E. OATES, *Fiscal Federalism* (1972); C.D. FOSTER, R.A. JACKMAN, and M. PERLMAN, *Local Government Finance in a Unitary State* (1980); and J. RICHARD ARONSON and JOHN HILLEY, *Financing State and Local Governments,* 4th ed. (1986).

The economic role of the government is analyzed in C.F. BASTABLE, *Public Finance,* 3rd rev. ed. (1903, reprinted 1917); JAMES M. BUCHANAN and MARILYN R. FLOWERS, *The Public Finances,* 5th ed. (1980); RICHARD STONE and GIOVANNA STONE, *National Income and Expenditure,* 10th ed. (1977); BENT HANSEN, *Fiscal Policy in Seven Countries, 1955–1965: Belgium, France, Germany, Italy, Sweden, United Kingdom, United States* (1969); CHARLES J. HITCH and ROLAND N. MCKEAN, *The Economics of Defense in the Nuclear Age* (1960, reissued 1975); JOHN KENNETH GALBRAITH, *The Affluent Society,* 4th ed. (1984); FRANCIS M. BATOR, *The Question of Government Spending: Public Needs and Private Wants* (1960); LEIF JOHANSEN, *Public Economics* (1965, reissued 1975; originally published in Norwegian, 1965); ERIK LUNDBERG, *Instability and Economic Growth* (1968); RONALD L. TEIGEN (ed.), *Readings in Money, National Income, and Stabilization Policy,* 4th ed. (1978); C.A.E. GOODHART, *Money, Information, and Uncertainty* (1975); JOHN G. GURLEY and EDWARD S. SHAW, *Money in a Theory of Finance* (1960, reprinted 1971); VICTORIA CHICK, *The Theory of Monetary Policy,* rev. ed. (1977); JOSEPH A. PECHMAN and BENJAMIN A. OKNER, *Who Bears the Tax Burden?* (1974); and ROBERT H. HAVEMAN and JULIUS MARGOLIS (eds.), *Public Expenditure and Policy Analysis,* 3rd ed. (1983).

(K.E.P./J.F.D./A.Li./C.N.M./Ed.)

Gravitation

Gravitation is a universal force of attraction acting between all matter. It is by far the weakest known force in nature and thus plays no role in determining the internal properties of everyday matter. Due to its long reach and universality, however, gravity shapes the structure and evolution of stars, galaxies, and the entire universe. The trajectories of bodies in the solar system are determined by the laws of gravity, while on Earth all bodies have a weight, or downward force of gravity, proportional to their mass, which the Earth's mass exerts on them. Gravity is measured by the acceleration that it gives to freely falling objects. At the Earth's surface, the acceleration of gravity is about 9.8 metres (32 feet) per second per second. Thus, for every second an object is in free fall, its speed increases by about 9.8 metres per second.

The works of Isaac Newton and Albert Einstein dominate the development of gravitational theory. Newton's classical theory of gravitational force held sway from his *Principia,* published in 1687, until Einstein's work in the early 20th century. Even today, Newton's theory is of sufficient accuracy for all but the most precise applications. Einstein's modern field theory of general relativity predicts only minute quantitative differences from the Newtonian theory except in a few special cases. The major significance of Einstein's theory is its radical conceptual departure from classical theory and its implications for further growth in physical thought.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 131 and 212.

This article is divided into the following sections:

Development of gravitational theory 169
- Early concepts
- Newton's law of gravity
- Potential theory

Acceleration around the Earth, Moon, and planets 171
- Variations in g
- Measurements of g
- Geophysics
- The Moon and the planets

Gravitational theory and other aspects of physical theory 173
- Field theories of gravitation
- Gravitational fields and general theory of relativity
- The paths of particles and light
- Gravitational radiation

Some astronomical aspects of gravitation 174

Experimental study of gravitation 175
- The inverse square law
- The principle of equivalence
- The constant of gravitation
- The variation of the constant of gravitation with time
- Fundamental character of G

Bibliography 177

DEVELOPMENT OF GRAVITATIONAL THEORY

Early concepts. Until Newton's findings, it was not realized that both the movement of celestial bodies and the free fall of objects on Earth are determined by the same force. The classical Greek philosophers, for example, did not consider the celestial bodies to be affected by gravity because the bodies were observed to follow perpetually repeating, non-descending trajectories in the sky. Thus,

Aristotle's views

Aristotle considered that each heavenly body followed a particular "natural" motion, unaffected by external causes or agents. Aristotle also believed that massive earthly objects possess a natural tendency to move toward the Earth's centre. These ideas and two other Aristotelian viewpoints prevailed for centuries: that a body moving at constant speed requires a continuous force acting on it and that force must be applied by contact rather than interaction at a distance. These views impeded the understanding of the principles of motion and precluded the development of ideas about universal gravitation. During the 16th and early 17th centuries, however, several scientific contributions to the problem of earthly and celestial motion set the stage for Newton's later gravitational theory.

Kepler's laws of planetary motion

The 17th-century German astronomer Johannes Kepler accepted the Copernican perspective in which the planets orbit the Sun rather than the Earth. Using the improved measurements of planetary movements made by the Danish astronomer Tycho Brahe during the 16th century, Kepler described the planetary orbits with simple geometric and arithmetic relations. Kepler's three quantitative laws of planetary motion are: (1) the planets describe elliptic orbits, of which the Sun occupies one focus (a focus is one of two points inside an ellipse; rays coming from one of the points bounce off any side of the ellipse and go through the other focus); (2) the line joining a planet to the Sun sweeps out equal areas in equal time; and (3) the square of the period of revolution of a planet is proportional to the cube of its average distance from the Sun. During this same period the Italian astronomer and physicist Galileo made progress in understanding "natural" motion and simple accelerated motion for earthly objects. He realized that bodies that are uninfluenced by forces continue indefinitely to move and that force is necessary to change motion, not to maintain constant motion. In studying how objects fall toward the Earth, Galileo discovered that the motion is one of constant acceleration. He was able to show that the distance a falling body travels from rest in this way varies as the square of the time. As noted above, the acceleration due to gravity at the surface of the Earth is about 9.8 metres per second per second.

Newton's law of gravity. Newton discovered the relationship between the motion of the Moon and the motion of any falling body on Earth. His gravitational theory explained Kepler's laws and established the modern quantitative science of gravitation. Newton assumed the presence of an attractive force between all massive bodies, one which does not require bodily contact and acts at a distance. By invoking his law of inertia (bodies not acted upon by a force move at constant speed in a straight line), Newton concluded that a force exerted by the Earth on the Moon is needed to keep it in a circular motion about the Earth rather than moving in a straight line. He realized that this force could be, at long range, the same as the force with which the Earth pulls objects on its surface downward. When Newton discovered that the acceleration of the Moon is $1/3{,}600$ smaller than the acceleration at the surface of the Earth, he related the number 3,600 to the radius of the Earth squared. He calculated that the circular orbital motion of radius R and period T requires a constant inward acceleration A equal to the product of $4\pi^2$ and the ratio of the radius to the square of the time:

$$A = \frac{4\pi^2 R}{T^2}. \qquad (1)$$

Newton applied his results to the Moon's orbit, which has a period of 27.3 days and a radius of about 384,000 kilometres (approximately 60 Earth radii). He found the Moon's inward acceleration in its orbit to be 0.0027 metre per second per second, the same as $(1/60)^2$ of the acceleration of a falling object at the surface of the Earth. Because Newton could thus relate the two accelerations to

a common interaction, he deduced that the gravitational force between bodies diminishes as the inverse square of the distance between the bodies. Thus, if the distance between the bodies is doubled, the force on them is reduced to one-quarter. A further result, that the mass of the Earth acts gravitationally on the outside world as if the mass were concentrated at the planet's centre, was needed to obtain his relationship. Newton proved it to be true for all spherically symmetric bodies.

Newton saw that the gravitational force between bodies must depend on the masses of the bodies. Since a body of mass M experiencing a force F accelerates at a rate F/M, a force of gravity proportional to M would be consistent with Galileo's observation that all bodies accelerate under gravity toward the Earth at the same rate. In Newton's equation, F_{12} is the magnitude of the gravitational force acting between masses M_1 and M_2 separated by distance r_{12}. The force equals the product of these masses and of G, a universal constant, divided by the square of the distance.

The universal constant G

$$F_{12} = \frac{GM_1M_2}{r_{12}^2} \tag{2}$$

The constant G is a quantity with the physical dimensions (length)^3/(mass)(time)^2; its numerical value depends on the physical units of length, mass, and time used. (G is discussed more fully in subsequent sections.)

The force acts in the direction of the line joining the two bodies and so is represented naturally as a vector, $\boldsymbol{F}$. If $\boldsymbol{r}$ is the vector separation of the bodies, then

$$\boldsymbol{F} = \frac{GM_1M_2\boldsymbol{r}}{r^3}. \tag{3}$$

In this expression, the factor $\boldsymbol{r}/r^3$ acts in the direction of $\boldsymbol{r}$ and is numerically equal to $1/r^2$.

The attractive force of a number of bodies of masses M_1 on a body of mass M is

$$\boldsymbol{F} = \frac{GM\Sigma_i M_i\boldsymbol{r}_i}{r_i^3}, \tag{4}$$

where Σ_i means that the vector forces due to all the attracting bodies must be added together. This is Newton's gravitational law essentially in its original form. A simpler expression, equation (5), gives the surface acceleration on Earth. Setting a mass equal to the Earth's mass M_E and the distance equal to the Earth's radius r_E, the downward acceleration of a body at the surface g is equal to the product of the universal gravitational constant and the mass of the Earth divided by the square of the radius:

$$g = \frac{GM_E}{r_E^2}. \tag{5}$$

Weight and mass

The weight W of the body can be measured by the equal and opposite force necessary to prevent the downward acceleration; this is M_g. The same body placed on the surface of the Moon has the same mass, but, as the Moon has a mass of about 1/81 times that of the Earth and a radius of just 0.27 that of the Earth, the body on the lunar surface has a weight of only 1/6 its Earth weight, as the U.S. Apollo astronauts demonstrated. Passengers and instruments in orbiting satellites, where no force prevents the free fall of the satellites in the gravitational field, experience weightless conditions even though their masses remain the same as on Earth.

Equations (1) and (2) above can be used to derive Kepler's third law for the case of circular planetary orbits. By using the expression for the acceleration A in equation (1) for the force of gravity for the planet GM_pM_S/R^2 divided by the planet's mass M_p, the following equation, in which M_S is the mass of the Sun, is obtained:

$$\frac{GM_S}{R^2} = \frac{4\pi^2R}{T^2}$$

or

$$R^3 = \left(\frac{GM_S}{4\pi^2}\right)T^2. \tag{6}$$

Newton was able to show that all three of Kepler's observationally derived laws follow mathematically from the assumption of his own laws of motion and gravity. In all observations of the motion of a celestial body, only the product of G and the mass can be found. Newton first estimated the magnitude of G by assuming the Earth's average mass density to be about 5.5 times that of water (somewhat greater than the Earth's surface rock density) and by calculating the Earth's mass from this. Then, taking M_E and r_E as the Earth's mass and radius, respectively, the value of G was

$$G = \frac{gr_E^2}{M_E}, \tag{7}$$

which numerically comes close to the accepted value of 6.6726×10^{-11} m^{3}s^{-2}/kg^{-1}, first directly measured by a Cavendish balance experiment (see below).

Comparing equation (5) above for the Earth's surface acceleration g with the R^3/T^2 ratio for the planets, a formula for the ratio of the Sun's mass, M_S, to Earth's mass, M_E, was obtained in terms of known quantities, R_E being the radius of the Earth's orbit:

$$\frac{M_S}{M_E} = \frac{4\pi^2R_E^3}{gT_E^2r_E^2} \cong 325{,}000. \tag{8}$$

Using observations of the motion of the moons of Jupiter discovered by Galileo, Newton determined that Jupiter was 318 times more massive than the Earth but only 1/4 as dense, having a radius 11 times larger than the Earth's.

Interaction among celestial bodies

When two celestial bodies of comparable mass interact gravitationally, both orbit about a fixed point (the centre of mass of the two bodies). This point lies between the bodies on the line joining them at a position such that the distances to each body multiplied by each body's mass are equal. Thus, the Earth and the Moon move in an orbit about their common centre of mass. This motion of the Earth has two observable consequences. First, the direction of the Sun as seen from the Earth relative to the very distant stars varies each month by about 12 arc seconds in addition to the Sun's annual motion. Second, the line-of-sight velocity from the Earth to a freely moving spacecraft varies each month by 2.04 metres per second according to very accurate data obtained from radio tracking. From these results the Moon is found to have a mass 1/81 times that of the Earth. With slight modifications Kepler's laws remain valid for systems of two comparable masses; the foci of the elliptical orbits are the two-body centre-of-mass positions, and, putting $M_1 + M_2$ instead of M_S in the expression of Kepler's third law, equation (6) above, the third law reads:

$$R^3 = \frac{G(M_1 + M_2)}{4\pi^2}T^2. \tag{9}$$

This agrees with equation (6) when one body is so small that its mass can be neglected. The rescaled formula can be used to determine the separate masses of binary stars (pairs of stars orbiting around each other) that are a known distance from the solar system. Equation (9) determines the sum of the masses; and, if R_1 and R_2 are the distances of the individual stars from the centre of mass, the ratio of the distances must balance the inverse ratio of the masses, and the sum of the distances is the total distance R. In symbols,

$$\frac{R_1}{R_2} = \frac{M_2}{M_1};\ R_1 + R_2 = R. \tag{10}$$

These relations are sufficient to determine the individual masses. Observations of the orbital motion of double stars, of the dynamic motion of stars collectively moving within their galaxies, and of the motion of the galaxies themselves verify that Newton's law of gravity is valid to a high degree of accuracy throughout the visible universe.

Ocean tides, phenomena that mystified thinkers for centuries, were also shown by Newton to be a consequence of the universal law of gravitation, although the details of the complicated phenomena were not understood until comparatively recently. They are caused specifically by the gravitational pull of the Moon and, to a lesser extent, of the Sun (see OCEANS: *Tide-generating forces*).

It was already known in Newton's day that the Moon does not move in a simple Keplerian orbit. Later, more accurate observations of the planets also showed discrep-

ancies from Kepler's laws. The motion of the Moon is particularly complex; however, apart from a long-term acceleration due to tides on the Earth, it can be accounted for by the gravitational attraction of the Sun, the Earth, and the other planets. In addition, the gravitational attraction of the planets for each other explains almost all the features of their motions. The exceptions are nonetheless important. Uranus, the seventh planet from the Sun, was observed to undergo variations in its motion that could not be explained by perturbations from Saturn, Jupiter, and the other planets. Two 19th-century astronomers, John Couch Adams of Britain and Urbain-Jean-Joseph Le Verrier of France, independently assumed the presence of an unseen eighth planet that could produce the observed discrepancies. They calculated its position within a degree of which the planet Neptune was discovered in 1846. Measurements of the motion of the innermost planet, Mercury, over an extended period led astronomers to conclude that the major axis of this planet's elliptical orbit precesses in space at a rate 43 arc seconds per century faster than could be accounted for from perturbations of the other planets. In this case, however, no other bodies could be found that could produce this discrepancy, and very slight modification of Newton's law of gravitation seemed to be needed. Einstein's theory of relativity precisely predicts this observed behaviour of Mercury's orbit (see RELATIVITY).

The discovery of Neptune

Potential theory. For irregular, nonspherical mass distributions in three dimensions, Newton's original vector equation (4, above) is inefficient, though theoretically it could be used for finding the resulting gravitational field. The main progress in classical gravitational theory after Newton was the introduction of potential theory, which is the mathematical representation of gravitational fields. It allows practical as well as theoretical investigation of the gravitational variations in space and of the anomalies due to the irregularities and shape deformations of the Earth.

Potential theory led to the following elegant formulation: the gravitational acceleration, g, a function of position $\boldsymbol{R}$, $\boldsymbol{g}(\boldsymbol{R})$, at any point in space is given from a function, Φ, called the gravitational potential, by means of a generalization of the operation of differentiation:

$$\boldsymbol{g}(\boldsymbol{R}) = \frac{\partial \Phi}{\partial x}\boldsymbol{i} + \frac{\partial \Phi}{\partial y}\boldsymbol{j} + \frac{\partial \Phi}{\partial z}\boldsymbol{k}, \qquad (11)$$

in which $\boldsymbol{i}$, $\boldsymbol{j}$, and $\boldsymbol{k}$ stand for unit basis vectors in a three-dimensional Cartesian coordinate system. The potential and therefore g are determined by an equation discovered by the French mathematician Siméon-Denis Poisson:

Poisson's equation

$$\left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} + \frac{\partial^2}{\partial z^2}\right)\Phi(\boldsymbol{R}) = -4\pi G\rho(\boldsymbol{R}). \qquad (12)$$

The significance of this approach is that Poisson's equation can be solved under rather general conditions, which is not the case with Newton's equation. When mass density ρ is nonzero, the solution is expressed as a definite integral:

$$\Phi(\boldsymbol{R}) = G\int \frac{\rho(\boldsymbol{R}')d\boldsymbol{R}'}{|\boldsymbol{R} - \boldsymbol{R}'|}. \qquad (13)$$

When $\rho = 0$ (in particular, outside the Earth), Poisson's equation reduces to the simpler equation of Laplace.

The appropriate coordinates for the region outside the nearly spherical Earth are spherical polar coordinates: $\boldsymbol{R}$, the distance from the centre of the Earth; θ, the colatitude measured from the North Pole; and the longitude measured from Greenwich. The solutions are series of powers of trigonometric functions of colatitude and longitude, known as spherical harmonics; the first terms are:

$$\Phi(\boldsymbol{R}) = \frac{GM_E}{R}\left[1 - J_2\left(\frac{R_E}{R}\right)^2 \frac{3\cos^2\theta - 1}{2}\right. \qquad (14)$$

$$\left. - J_3\left(\frac{R_E}{R}\right)^3 \frac{5\cos^3\theta - 3\cos\theta}{2} + \ldots\right],$$

The constants J_2, J_3, and so forth are determined by the detailed mass distribution of the Earth; and, since Newton showed that for a spherical body all the J_n are zero, they must measure the deformation of the Earth from a spherical shape. J_2 measures the magnitude of the Earth's rotational equatorial bulge, J_3 measures a slight pear-shaped deformation of the Earth, and so on. From observations of perturbations on satellite orbits, the parameters J_2 and J_3 have been found to be $1{,}082.7 \times 10^{-6}$ and -2.4×10^{-6}, respectively.

Much the largest variation in the potential around the Earth is due to the equatorial bulge, and there is a corresponding increase in the value of gravity on the surface of the Earth from the Equator to the poles. Newton was the first to give a theory of the equatorial bulge; among the data he used to estimate its size were some pendulum measurements made at Saint Helena by the English astronomer Edmond Halley (see below). Many more harmonic components have been estimated from observations of the orbits of artificial satellites close to Earth; the sum of the most important of them can be shown by contours on a map of Earth's surface at sea level, where the potential is a constant (see Figure 1).

From Alan Cook, *Physics of the Earth and Planets* (1973); Macmillan Publishers Ltd., London

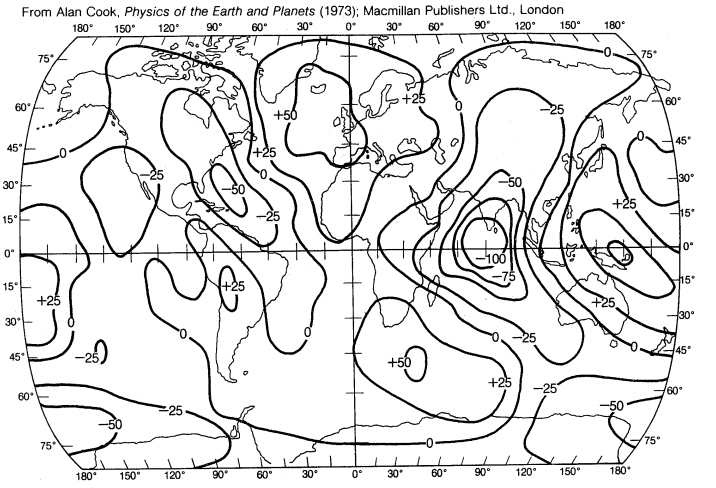


Figure 1: The gravity field of the Earth as indicated by the shape of the sea-level surface; contours (in metres) of deviations from a spheroid of flattening 1/298.25.

Effects of local mass differences. Spherical harmonics are the natural way of expressing the large-scale variations of potential that arise from the deep structure of the Earth. However, spherical harmonics are not suitable for local variations due to more superficial structures. Not long after Newton's time, it was found that the gravity on top of large mountains is less than expected on the basis of their visible mass. The idea of isostasy was developed, according to which the unexpectedly low acceleration of gravity on a mountain is caused by low-density rock 30 to 100 kilometres underground, which buoys up the mountain. Correspondingly, the unexpectedly high force of gravity on ocean surfaces is explained by dense rock 30 to 100 kilometres beneath the ocean bottom.

Concept of isostasy

Portable gravimeters, which can detect variations of one part in 10^9 in the gravitational force, are in wide use today for mineral and oil prospecting. Unusual underground deposits reveal their presence by producing local gravitational variations (see below).

Weighing the Earth. Because the mass of the Earth can be calculated from g and the planet's radius if G is known, the English physicist-chemist Henry Cavendish and other early experimenters termed their work "weighing the Earth." The mass of the Earth is about 5.98×10^{24} kilograms, while the mean densities of the Earth, Sun, and Moon are, respectively, 5.52, 1.43, and 3.3 times that of water. (K.L.N./A.H.C.)

ACCELERATION AROUND THE EARTH, MOON, AND PLANETS

The value of the attraction of gravity or of the potential is determined by the distribution of matter within the Earth or other celestial body. In turn, the distribution of matter determines the shape of the surface on which the potential is constant (Figure 1). Measurements of gravity and the potential are thus essential both to geodesy, which is the study of the shape of the Earth, and to geophysics, the

study of its internal structure. For geodesy and global geophysics, it is best to measure the potential from the orbits of artificial satellites. Surface measurements of gravity are best for local geophysics, which deals with the structure of mountains and oceans and the search for minerals.

Variations in g. *Changes due to location.* The acceleration g varies by about $1/2$ of 1 percent with position on the Earth's surface, from about 9.78 metres per second per second at the Equator to approximately 9.83 metres per second per second at the poles. This variation stems chiefly from the rotation of the Earth, as part of the Earth's pull is balanced by keeping objects rotating with the Earth instead of flying tangentially off into space (as mud does off a spinning wheel). This effect is also responsible for the bulge of the Earth at the Equator and the slight flattening at the poles. The distance to the centre of the Earth, therefore, increases with the bulge from the poles to the Equator, and consequently the value of g, which for a spherical Earth of radius r_E and mass M_E is given simply by GM_E/r_E^2, is less toward the Equator.

In addition to this broad-scale variation, local variations of a few parts in 10^6 or smaller are caused by variations in the density of the Earth's crust as well as height above sea level.

Changes with time. At any one place, g varies with time as the result of the changing positions of the Sun and the Moon—namely, the tidal variation. For most purposes it is necessary to know only the variation of gravity with time at a fixed place or the changes of gravity from place to place; then, the tidal variation can be removed. Accordingly, almost all gravity measurements are relative measurements of the differences from place to place or from time to time.

Measurements of g. *Unit of gravity.* Because gravity changes are far less than one metre per second per second, it is convenient to have a smaller unit for relative measurements. The gal (after Galileo) has been adopted for this purpose; one gal equals $1/100$ of one metre per second per second. The unit most commonly used is the milligal (mgal), which equals 10^{-5} metre per second per second—*i.e.*, $1/1{,}000{,}000$ of the average value of g.

Absolute measurements. Two basic ways of making absolute measurements of gravity have been devised: timing the free-fall of an object and timing the motion under gravity of a body constrained in some way, almost always by a pendulum. In 1817 the English physicist Henry Kater, building on the work of the German astronomer Friedrich Wilhelm Bessel, was the first to use a reversible pendulum to make absolute measurements of g. If the periods of swing of a rigid pendulum about two alternative points of support are the same, then the separation of those two points is equal to the length of the equivalent simple pendulum of the same period. By careful construction, Kater was able to measure the separation very accurately. The so-called reversible pendulum was used for absolute measurements of gravity from Kater's day until the 1950s. Since that time, electronic instruments have enabled investigators to measure with high precision the half-second time of free-fall of a body (from rest) through one metre. It is also possible to make extremely accurate measurements using light interference. Consequently, direct measurements of free-fall have replaced the pendulum for absolute measurements of gravity. Nowadays lasers serve as light sources for interferometers; the falling object, in effect, becomes a retroreflector that returns a beam of light back upon itself. Transportable versions of such apparatus have been used in different locations to establish a basis for measuring differences of gravity over the entire Earth. The accuracy attainable in these measurements is about one part in 10^8.

The work of Henry Kater

Relative measurements. From the time of Newton, measurements of differences of gravity (strictly, the ratios of values of gravity) were made by timing the same pendulum at different places. During the 1930s, however, static gravimeters replaced pendulums for local measurements over small ranges of gravity. Today, free-fall measurements have rendered the pendulum obsolete for all purposes.

Spring gravimeters balance the force of gravity mg on a mass m in the gravity field to be measured, against the elastic force of the spring, using either electronic or mechanical means to achieve high sensitivity. Vibrating string gravimeters in which the string's vibration frequency is determined by g also have been developed. A device of this type was employed by the Apollo 17 astronauts on the Moon to conduct a gravity survey of their landing site. Another relatively recent development is the superconducting gravimeter, an instrument in which the position of a magnetically levitated superconducting sphere is sensed to provide a measure of g.

Spring gravity meters

Modern gravimeters may have sensitivities better than 0.005 milligal, the standard deviation of observations in exploration surveys being on the order of 0.01–0.02 milligal.

Differences in gravity measured with gravimeters are obtained in quite arbitrary units—divisions on a graduated dial, for example. The relation between these units and milligals can only be determined by reading the instrument at a number of points where g is known as a result of absolute or relative pendulum measurements. Further, because an instrument will not have a completely linear response, known points must cover the entire range of gravity over which the gravimeter is to be used.

Gravimetric surveys. Recently, by combining all available absolute and relative measurements, it has been possible to obtain the most probable gravity values at a large number of sites to a high degree of accuracy. The culmination of gravimetric work begun in the 1960s has been a worldwide gravity reference system having an accuracy of at least one part in 10^7 (0.1 milligal or better).

Since g is an acceleration, the problem of its measurement from a vehicle that is moving and therefore unavoidably accelerating relative to the Earth raises a number of fundamental problems. Pendulum, vibrating-string, and spring-gravimeter observations have been made from submarines; using gyrostabilized platforms, relative gravity measurements with accuracies approaching a few milligals have been and are being made from surface ships. Experimental measurements with various gravity sensors on fixed-wing aircraft as well as on helicopters have been carried out.

Geophysics. The value of gravity measured at the terrestrial surface is the result of (1) the gravitational attraction of the Earth as a whole, (2) centrifugal force caused by the Earth's rotation, (3) elevation, (4) unbalanced attractions caused by surface topography, (5) tidal variations, and (6) unbalanced attractions caused by irregularities in underground density distributions. Most geophysical surveys are aimed at separating out the last of these in order to interpret the geologic structure. It is therefore necessary to make proper allowance for the other factors.

The first two factors imply a variation of gravity with latitude that can be calculated for an assumed shape for the Earth. The third factor, which is the decrease in gravity with elevation, due to increased distance from the centre of the Earth, amounts to -0.3086 milligal per metre. This value, however, assumes that material of zero density occupies the whole space between the point of observation and sea level, and it is therefore termed the free-air correction factor. In practice, the mass of rock material that occupies part or all of this space must be considered. In an area where the topography is reasonably flat, this is usually calculated by assuming the presence of an infinite slab of thickness equal to the height of the station, h, and having an appropriate density σ; its value is $+0.04185\ \sigma h$ milligal per metre. This is commonly called the Bouguer correction factor.

The free-air and Bouguer corrections

Terrain or topographic corrections also can be applied to allow for the attractions due to surface relief if the densities of surface rocks are known. Tidal effects (the amplitudes are less than 0.3 milligal) can be calculated and allowed for. (Ja.F./A.H.C.)

The Moon and the planets. It was noted above that the large-scale features of the Earth's potential are obtained from observations of Earth-orbiting satellites and may be shown as a map of the sea-level surface (Figure 1). Although the Apollo astronauts used a gravimeter at their lunar landing site, most scientific knowledge about the Moon's gravitational attraction has been derived from observations of its effects on the accelerations of spacecraft. Information about the gravitational force of various plan-

ets has been obtained in this way as well. Radio tracking makes it possible to estimate the accelerations of spacecraft very accurately, and the results can be expressed either as terms in a series of spherical harmonics or as the variation of gravity over the surface. As in the case of the Earth, spherical harmonics are more effective for studying gross structure, while the variation of gravity is more useful for local features. Because spacecraft must descend close to the surface to detect local gravity variations, only data for the Moon and Mars have been obtained so far.

The Moon's polar flattening is much less than that of the Earth, while its equator is far more elliptical. There are also large, more local irregularities from visible and concealed structures. Mars also exhibits some large local variations, while the equatorial bulges of Mercury and Venus are very slight. By contrast, the major planets, all of which rotate quite fast, have large equatorial bulges, and their gravity is dominated by a large increase from equator to pole. These results are crucial to understanding the internal properties of the planets.

GRAVITATIONAL THEORY AND OTHER ASPECTS OF PHYSICAL THEORY

The Newtonian theory of gravity is based on an assumed force acting between all pairs of bodies—*i.e.,* an action at a distance. When a mass moves, the force acting on other masses has been considered to adjust instantaneously to the new location of the displaced mass. Special relativity theory states that no physical signal travels faster than the speed of light and that all signals travel at this speed through empty space. This theory, with the field theory of electrical and magnetic phenomena, has met such empirical success that most modern gravitational theories are constructed as field theories consistent with the principles of special relativity. In a field theory the gravitational force between bodies is formed by a two-step process: (1) One body produces a gravitational field that permeates all surrounding space but has weaker strength farther from its source. A second body in that space is then acted upon by this field and experiences a force. (2) The Newtonian force of reaction is then viewed as the response of the first body to the gravitational field produced by the second body, there being at all points in space a superposition of gravitational fields due to all the bodies in it.

Special relativity and electromagnetic field theory

Field theories of gravitation. The possibility that gravitation might be linked with the other forces of nature in a unified theory of forces greatly increased interest in gravitational field theories during the 1970s and '80s. The first such unified field theory, and so far the only successful one, is that of physicists Abdus Salam of Pakistan and Steven Weinberg and Sheldon L. Glashow of the United States, who proposed that the electromagnetic forces and the weak force responsible for beta decay are different manifestations of one and the same basic interaction. Physicists now are actively seeking other possible unified combinations. Because the gravitational force is exceedingly weak compared with all others and because it seems to be independent of all physical properties except mass, the unification of gravitation with the other forces will be the most difficult to achieve. This challenge has provided a tremendous impetus to experimental investigations to determine whether there may be some failure of the apparent independence.

The prime example of a field theory is Einstein's general relativity, according to which the acceleration due to gravity is a purely geometric consequence of the properties of space-time in the neighbourhood of attracting masses. (As will be seen below, general relativity makes certain specific predictions that are borne out well by observation.) In a whole class of more general theories, these and other effects not predicted by simple Newtonian theory are characterized by free parameters; such formulations are called parameterized post-Newtonian (PPN) theories. There is now considerable experimental and observational evidence for limits to the parameters. So far, no deviation from general relativity has been demonstrated convincingly.

Field theories of gravity predict specific corrections to the Newtonian force law, the corrections taking two basic forms: (1) When matter is in motion, additional gravitational fields (analogous to the magnetic fields produced by moving electric charges) are produced; also, moving bodies interact with gravitational fields in a motion-dependent way. (2) Unlike electromagnetic field theory, in which two or more electric or magnetic fields superimpose by simple addition to give the total fields, in gravitational field theory nonlinear fields proportional to the second and higher power of the source masses are generated, and gravitational fields proportional to the product of different masses are created. Gravitational fields themselves become sources for additional gravitational fields. Examples of some of these effects are shown below. The acceleration, A, of a moving particle of negligible mass that interacts with a mass, M, which is at rest, is given in the following formula, derived from Einstein's gravitational theory. The expression for A now has, as well as the Newtonian expression from equation (1), further terms in higher powers of GM/R^2—that is, in G^2M^2/R^4. As elsewhere, $\boldsymbol{V}$ is the particle's velocity vector, $\boldsymbol{A}$ is its acceleration vector, $\boldsymbol{R}$ is the vector from the mass M, and c is the speed of light. When written out, the sum is

Corrections of Newtonian force law

$$A = -\frac{GM\boldsymbol{R}}{R^3} + 2\frac{G^2M^2\boldsymbol{R}}{c^2R^2} - \frac{3}{2}\frac{GM\boldsymbol{R}}{R^3}\left(\frac{V^2}{c^2}\right) - \frac{\boldsymbol{V}\cdot\boldsymbol{A}\boldsymbol{V}}{c^2} - \frac{1}{2}\frac{\boldsymbol{V}^2}{c^2}\boldsymbol{A} + \ldots . \tag{15}$$

This expression gives only the first post-Newtonian corrections; terms of higher power in $1/c$ are neglected. For planetary motion in the solar system the $1/c^2$ terms are smaller than Newton's acceleration term by at least a factor of 10^{-8}, but some of the consequences of these correction terms are measurable and important tests of Einstein's theory. It should be pointed out that prediction of new observable gravitational effects requires particular care; Einstein's pioneer work in gravity has shown that gravitational fields affect the basic measuring instruments of experimental physics—clocks, rulers, light rays—with which any experimental result in physics is established. Some of these effects are listed below:

1. The rate at which clocks run is reduced by proximity of massive bodies; *i.e.,* clocks near the Sun will run slowly compared with identical clocks farther away from it.
2. In the presence of gravitational fields the spatial structure of physical objects is no longer describable precisely by Euclidean geometry; for example, in the arrangement of three rigid rulers to form a triangle, the sum of the subtended angles will not equal 180°. A more general type of geometry, Riemannian geometry, seems required to describe the spatial structure of matter in the presence of gravitational fields.
3. Light rays do not travel in straight lines, the rays being deflected by gravitational fields. To distant observers the light-propagation speed is observed to be reduced near massive bodies.

Gravitational fields and general theory of relativity. In Einstein's general theory of relativity the physical consequences of gravitational fields are stated in the following way. Space-time is a four-dimensional non-Euclidean continuum, the curvature of space-time's Riemannian geometry being produced by or related to the world's matter distribution. Particles and light rays travel along the geodesics (shortest paths) of this four-dimensional geometric world.

There are two principal consequences of the geometric view of gravitation: (1) the accelerations of bodies depend only on their masses and not on their chemical or nuclear constitution, and (2) the path of a body or of light in the neighbourhood of a massive body (the Sun, for example) is slightly different from that predicted by Newton's theory. The first is the weak principle of equivalence. Newton himself performed experiments with pendulums that demonstrated the principle to better than one part in 1,000 for a variety of materials, and, at the beginning of the 20th century, the Hungarian physicist Roland, Baron von Eötvös, showed that different materials accelerate in the Earth's field at the same rate to within one part in 10^9. More recent experiments have shown the equality of

Weak principle of equivalence

accelerations in the field of the Sun to within one part in 10^{11}. Newtonian theory is in accord with these results because of the postulate that gravitational force is proportional to a body's mass.

Inertial mass is a mass parameter giving the inertial resistance to acceleration of the body when responding to all types of force. Gravitational mass is determined by the strength of the gravitational force experienced by the body when in the gravitational field g. The Eötvös experiments, therefore, show that the ratio of gravitational and inertial mass is the same for different substances.

Einstein's special theory of relativity views inertial mass as a manifestation of all the forms of energy in a body according to his fundamental relationship $E = mc^2$, E being the total energy content of a body, m the inertial mass of the body, and c the speed of light. Dealing with gravitation, then, as a field phenomenon, the weak principle of equivalence indicates that all forms of nongravitational energy must identically couple to or interact with the gravitational field, because the various materials in nature possess different fractional amounts of nuclear, electrical, magnetic, and kinetic energies, yet they accelerate at identical rates.

In the general theory of relativity the gravitational field also interacts with gravitational energy in the same manner as with other forms of energy, an example of that theory's universality not possessed by most other theories of gravitation.

The Sun has an appreciable fraction of internal gravitational energy, and the repetitions of the Eötvös experiments during the 1970s with the Sun instead of the Earth as the attracting mass revealed that bodies accelerate at identical rates in the Sun's field as well as in that of the Earth. Extremely accurate laser measurements of the distance of the Moon from the Earth have made possible a further test of the weak principle of equivalence. The chemical constitutions of the Earth and the Moon are not the same, and so, if the principle did not hold, they might accelerate at different rates under the Sun's attraction. No such effect has been detected.

Newton's third dynamic law states that every force implies an equal and opposite reaction force. Modern field theories of force contain this principle by requiring every entity that is acted upon by a field to be also a source of the field. An experiment by the American physicist Lloyd Kreuzer established to a one-part-in-20,000 accuracy that different materials produce gravitational fields with a strength the same as that of gravitational fields acting upon them. In this experiment a sphere of solid material was moved through a liquid of identical weight density. The absence of a gravitational effect on a nearby Cavendish balance instrument during the sphere's motion is interpreted as showing that the two materials had equal potency in producing a local gravitational-field anomaly.

The gravitational redshift

Other experiments have brought confirmation of Einstein's predictions to an accuracy of a few percent. Using the Mössbauer effect to monitor the nuclear reabsorption of resonant gamma radiation, a shift of wavelength of the radiation that traveled vertically tens of metres in the Earth's gravitational field was measured, and the slowing of clocks (in this case the nuclear vibrations are clocks) as predicted by Einstein was confirmed to 1 percent precision. If ν and $\Delta\nu$ are clock frequency and change of frequency, respectively, h is the height difference between clocks in the gravitational field g. Then

$$\frac{\Delta\nu}{\nu} = -\frac{gh}{c^2}. \tag{16}$$

For a height of 10 metres, there is only a one-part-in-10^{15} change in clock rates.

The paths of particles and light. The idea that light should be deflected by passing close to a massive body had been suggested by the British astronomer and geologist John Michell in the 18th century. However, Einstein's general relativity theory predicted twice as much deflection as Newtonian physics. Quick confirmation of Einstein's result came from measuring the direction of a star close to the Sun during an expedition led by the British astronomer Sir Arthur Stanley Eddington to observe the solar eclipse of 1919. Optical determinations of the change of direction of a star are subject to many systematic errors, and far better confirmation of Einstein's general relativity theory has been obtained from measurements of a closely related effect—namely, the increase of the time taken by electromagnetic radiation along a path close to a massive body. In fact, both effects may be understood as due to a decrease in the speed of light near a massive object. Timing the round-trip travel time for radar pulses between the Earth and other inner planets or artificial satellites passing behind the Sun, experiments have confirmed to about 4 percent the prediction of an additional time delay, Δt. This additional time delay is given by the following formula, in which M_S is the Sun's mass, R_1 and R_2 are the distances from the Sun to Earth and to the other reflecting body, and D is the distance of closest approach of the radar pulses to the Sun (ln stands for natural logarithm):

$$\Delta t = \frac{4GM_S}{c^3} \ln \frac{4R_1R_2}{D^2}. \tag{17}$$

The additional precession of the orbit of Mercury of 43 arc seconds per century (see above) was known before the development of the theory of general relativity. With radar measurements of the distances to the planets, similar anomalous precessions have been estimated for Venus and the Earth and have been found to agree with general relativity.

Gravitational radiation. An important consequence of field theories of gravitation is that the gravitational field itself can oscillate in the same way that an electromagnetic field can. Thus, if the masses that are the source of a field change with time, they may be expected to radiate energy through waves in the field. The radiation would be very weak, and, though some extremely sensitive instruments have been built in an effort to detect gravitational radiation, none has been observed with certainty so far. The principal sources are expected to be double stars in their regular motions and massive stars collapsing as supernovas. Even though no gravitational radiation has yet been detected on Earth, there are strong grounds for believing that it exists. One particular double star system has a pulsar as one of its components, and, from measurements of the Doppler shifts of the pulsar frequency, precise estimates of the period of the orbit show that the period is changing, corresponding to a decrease in the energy of the orbital motion. Gravitational radiation is the only known means by which this could happen.

Possible sources of gravitational radiation

(K.L.N./A.H.C.)

SOME ASTRONOMICAL ASPECTS OF GRAVITATION

As stated above, studies of gravity enable researchers to determine the masses and densities of celestial bodies and thereby make it possible to investigate the physical constitutions of stars and planets. Because gravitation is a very weak force, however, its distinctive effects appear only when masses are extremely large. The idea that light might be attracted gravitationally had been suggested by the aforementioned John Michell and examined by the French mathematician and astronomer Pierre-Simon Laplace. Predictions by classical physics and general relativity that light passing close to the Sun might be deflected are described above. There are two further consequences for astronomy. Light from a distant object may pass close to objects other than the Sun and be deflected by them. In particular, they may be deflected by a massive galaxy. If some object is behind a massive galaxy, as seen from Earth, deflected light may reach the Earth by more than one path. Operating like a lens that focuses light along different paths, the gravity of the galaxy may make the object appear multiple; examples of such apparently double objects have been found.

Both Michell and Laplace pointed out that the attraction of a very dense object upon light might be so great that the light could never escape from the object, rendering it invisible. Such a phenomenon is a black hole. The relativistic theory of black holes has been thoroughly developed in recent years, and astronomers have conducted an intense search for them. One possible class of black holes comprises stars that have used up all their nuclear

Black holes

energy so that they are no longer held up by radiation pressure and have collapsed to neutron stars of extremely high density. Black holes also are thought to be at the centres of the larger galaxies and to be the sources of the energy found there.

Black holes from which no radiation is able to escape cannot be seen by their own light, but there may be observable secondary effects. If a black hole were one component of a double star, the orbital motion of the pair and the mass of the invisible member might be derived from the oscillatory motion of a visible companion. Because black holes attract matter, any gas in the vicinity of an object of this kind would fall into it and acquire, before vanishing into the hole, a high velocity and consequently a high temperature. The gas may become hot enough to produce X rays and gamma rays from around the hole. While there is still no definite proof, such a mechanism may be the origin of at least some powerful X-ray and radio astronomical sources, including those at the centres of galaxies and quasars (see COSMOS: *Black-hole model for active galactic nuclei*). Only astronomical objects can be expected to be sources of detectable gravitational radiation. As already mentioned, gravitational radiation is probably responsible for secular changes in the orbits of some double stars, and so in the very long term it may have an effect on the stability of celestial objects. If and when gravitational radiation is detected, new astronomical phenomena will no doubt be discovered.

EXPERIMENTAL STUDY OF GRAVITATION

The essence of Newton's theory of gravitation is that the force between two bodies is proportional to the product of their masses and the inverse square of their separation and that the force depends on nothing else. With a small geometric modification, the same is true in general relativity. Newton himself tested his assumptions by experiment and observation. He made pendulum experiments to confirm the principle of equivalence and checked the inverse square law as applied to the periods and diameters of the orbits of the satellites of Jupiter and Saturn. During the latter part of the 19th century many experiments showed the force of gravity to be independent of temperature, electromagnetic fields, shielding by other matter, orientation of crystal axes, and other factors. The revival of such experiments during the 1970s was the result of theoretical attempts to relate gravitation to other forces of nature by showing that general relativity incompletely describes gravity. New experiments on the equivalence principle were performed, and experimental tests of the inverse square law were made in the laboratory. There also has been a continuing interest in the determination of the constant of gravitation, G, although it must be pointed out that G occupies a rather anomalous position among the other constants of physics. In the first place, the mass, M, of any celestial object cannot be determined independently of the gravitational attraction that it exerts. Thus, the combination GM, not the separate value of M, is the only meaningful property of a star, planet, or galaxy. Second, according to general relativity and the principle of equivalence, G does not depend on material properties but is in a sense a geometric factor. Hence, the determination of the constant of gravitation does not seem as essential as the measurement of quantities like the electronic charge or Planck's constant. It is also much less well determined experimentally than any of the other constants of physics.

The constant of gravitation as a geometric factor

Experiments on gravitation are in fact very difficult, as a comparison of experiments on the inverse square law of electrostatics with those on gravitation will show. The electrostatic law has been established to within one part in 10^{16} by using the fact that the field inside a closed conductor is zero when the inverse square law is in effect and then detecting any residual fields with very sensitive electronic devices. Gravitational forces have to be detected by mechanical means, most often the torsion balance, and, though the sensitivities of mechanical devices have been greatly improved, they are still far less sensitive than electronic equipment. Mechanical arrangements also preclude the use of a complete gravitational enclosure. Last, extraneous disturbances are relatively large because gravitational forces are very small (something that Newton first pointed out). Thus, the inverse square law is established over laboratory distances to no better than one part in 10^4.

The inverse square law. The current interest in the inverse square law arises from two suggestions. First, the gravitational field itself might have a mass, in which case the constant of gravitation would change in an exponential manner from one value for small distances to a different one for large distances over a characteristic distance related to the mass of the field. Second, the observed field might be the superposition of two or more fields of different origin and different strengths, one of which might depend on chemical or nuclear constitution. Deviations from the inverse square law have been sought in three ways: (1) the law has been checked in the laboratory over distances up to about one metre, (2) the effective value of G for distances between 100 metres and one kilometre has been estimated from geophysical studies, and (3) there have been careful comparisons of the value of the attraction of the Earth as measured on the surface and as experienced by artificial satellites.

Search for deviations from the inverse square law

Early in the 1970s an experiment by the American physicist Daniel R. Long seemed to show deviation from the inverse square law at a range of about 0.1 metre. Long compared the maximum attractions of two rings upon a test mass hung from the arm of a torsion balance. The maximum attraction of a ring occurs at a particular point on the axis and is determined by the mass and dimensions of the ring. If the ring is moved until the force on the test mass is greatest, the distance between the test mass and the ring is not needed. Two later experiments over the same range showed no deviation from the inverse square law. In one, conducted by the American physicist Riley Newman and his colleagues, a test mass hung on a torsion balance was moved around in a long hollow cylinder. The cylinder approximates a complete gravitational enclosure and, allowing for a small correction because it is open at the ends, the force on the test mass should not depend on its location within the cylinder. No deviation from the inverse square law was found. In the other experiment, performed in Cambridge, Eng., by Y.T. Chen and associates, the attractions of two solid cylinders of different mass were balanced against a cylinder with a third mass so that only the separations of the cylinders had to be known; it was not necessary to know the distances of any from a test mass. Again no deviation of more than one part in 10^4 from the inverse square law was found. Other somewhat less sensitive experiments at ranges up to one metre or so also have failed to establish any greater deviation.

The geophysical tests go back to a method for the determination of the constant of gravitation that had been used in the 19th century, especially by the British astronomer Sir George Airy. Suppose the value of gravity, g, is measured at the top and bottom of a horizontal slab of rock of thickness t and density d. The values for the top and bottom will be different for two reasons. First, the top of the slab is t farther from the centre of the Earth, and so the measured value of gravity will be less by $2(t/R)g$, where R is the radius of the Earth. Second, the slab itself attracts objects above and below it toward its centre; the difference between the downward and upward attractions of the slab is $4\pi Gtd$. Thus, a value of G may be estimated. Frank D. Stacey and his colleagues in Australia have made such measurements at the top and bottom of deep mine shafts and have claimed that there may be a real difference between their value of G and the best value from laboratory experiments. The difficulties lie in obtaining reliable samples of the density and in taking account of varying densities at greater depths.

Geophysical tests of the inverse square law

The absolute measurements of gravity described earlier, together with the comprehensive gravity surveys made over the surface of Earth, allow the mean value of gravity over the Earth to be estimated to about one part in 10^6. The techniques of space research also have given the mean value of the radius of the Earth and the distances of artificial satellites to the same precision; thus, it has been possible to compare the value of gravity on the Earth with that acting on an artificial satellite. Agreement to about one part in 10^6 shows that, over distances from the surface

of the Earth to close satellite orbits, the inverse square law is followed.

It seems at present that all the most reliable experiments and observations reveal no deviation from the inverse square law, although there should possibly be further investigations on the geophysical scale.

The principle of equivalence. Experiments with ordinary pendulums test the principle of equivalence to no better than about one part in 10^5. Eötvös obtained much better discrimination with a torsion balance. His tests depended on comparing gravitational forces with inertial forces for masses of different composition. Eötvös set up a torsion balance to compare, for each of two masses, the gravitational attraction of the Earth with the inertial forces due to the rotation of the Earth about its polar axis. His arrangement of the masses was not optimal, and he did not have the sensitive electronic means of control and reading that are now available. Nonetheless, Eötvös found that the weak equivalence principle (see above) was satisfied to within one part in 10^9 for a number of very different chemicals, some of which were quite exotic. His results were later confirmed by the Hungarian physicist J. Renner. Renner's work has been analyzed recently in great detail because of the suggestion that it could provide evidence for a new force. It seems that the uncertainties of the experiments hardly allow such analyses, however.

Eötvös suggested that the attraction of the Sun upon test masses could be compared with the inertial forces of the Earth's orbital motion about the Sun. He performed some experiments, verifying equivalence with an accuracy similar to that which he had obtained with his terrestrial experiments. The solar scheme has substantial experimental advantages, and the American physicist Robert H. Dicke and his colleagues, in a careful series of observations in the 1960s (employing up-to-date methods of servo control and observation), found that the weak equivalence principle held to about one part in 10^{11} for the attraction of the Sun on gold and aluminum (Figure 2). A later experiment, with very different experimental arrangements, by Vladimir Braginski of the Soviet Union, gave a limit of about one part in 10^{12} for platinum and aluminum.

From Alan Cook, "Experiments on Gravitation," *Reports on Progress in Physics*, 51(5):707–757, May 1988

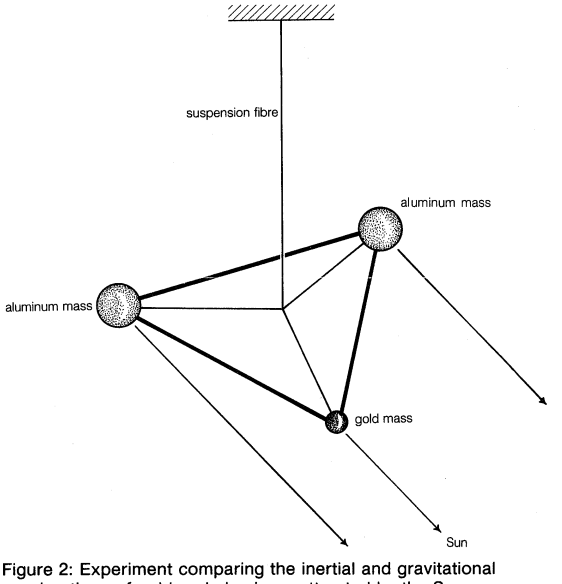


Figure 2: Experiment comparing the inertial and gravitational accelerations of gold and aluminum attracted by the Sun.

Measurements of the gravitational constant

The constant of gravitation. The constant of gravitation has been measured in three ways: (1) the comparison of the pull of a large natural mass with that of the Earth, (2) the measurement with a laboratory balance of the attraction of the Earth upon a test mass, and (3) the direct measurement of the force between two masses in the laboratory. The first approach was suggested by Newton; the earliest observations were made in 1774 by the British astronomer Nevil Maskelyne on the mountain of Schiehallion in Scotland. The subsequent work of Airy

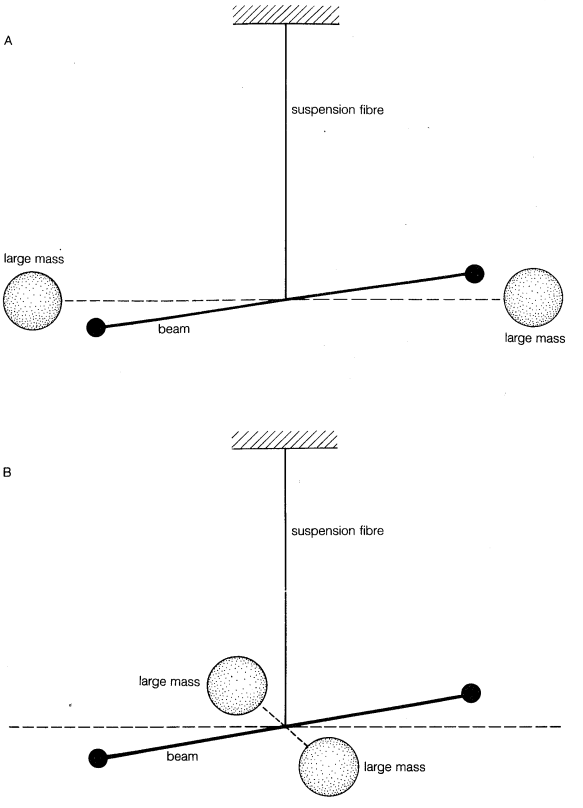


Figure 3: Principle of the determination of G from the change in period of a torsion balance with attracting masses in different positions.

From Alan Cook, "Experiments on Gravitation," *Reports on Progress in Physics*, 51(5):707–757, May 1988

and more recent developments are noted above. The laboratory balance method was developed in large part by the British physicist John Henry Poynting during the late 1800s, but all the most recent work has involved the use of the torsion balance in some form or other for the direct laboratory measurement of the force between two bodies. The torsion balance was devised by Michell, who died before he could use it to measure G. Cavendish adapted Michell's design to make the first reliable measurement of G in 1798; his results were so good in fact that it is only in comparatively recent times that clearly better results have been obtained. Cavendish himself measured the change in deflection of the balance when attracting masses were moved from one side to the other of the torsion beam. The method of deflection was analyzed most thoroughly in the late 1800s by Sir Charles Vernon Boys, an English physicist, who carried it to its highest development using a delicate suspension fibre of fused silica for the pendulum.

The second scheme involves observing the changes in the period of oscillation of a torsion balance when attracting masses are placed close to it such that the period is shortened in one position and lengthened in another (see Figure 3). Period measurements can be made much more precisely than those of deflection, and this scheme, introduced by Carl Braun of Austria in 1897, seems to have given the best results so far. Indeed it is only recently that the value obtained by the American physicist Paul R. Heyl and his colleagues has been improved upon by Gabriel G. Luther and William R. Towler (see table).

In the third scheme, a balance with heavy attracting masses is set up near a free test balance and adjusted so that it oscillates with the same period as the test balance. The latter is then driven into resonant oscillations with an amplitude that is a measure of the constant of gravitation. The technique was first employed by J. Zahradnicek of Czechoslovakia during the 1930s and was effectively used again by C. Pontikis of France some 40 years later.

The variation of the constant of gravitation with time. The 20th-century English physicist P.A.M. Dirac, among others, suggested that the value of the constant of gravi-

Values of the Constant of Gravitation (up to 1982)

author	year	method	G (in units of 10^{-11} $m^3s^{-2}kg^{-1}$)
H. Cavendish	1798	torsion-balance (deflection)	6.754
J.H. Poynting	1891	common-balance	6.698
C.V. Boys	1895	torsion-balance (deflection)	6.658
C. Braun	1897	torsion-balance (deflection)	6.658
	1897	torsion-balance (period)	6.658
P.R. Heyl	1930	torsion-balance (period)	6.669
J. Zahradnicek	1932	torsion-balance (resonance)	6.659
P.R. Heyl, P. Chrzanowski	1942	torsion-balance (period)	6.672
C. Pontikis	1972	torsion-balance (resonance)	6.6714
G.G. Luther and W.R. Towler	1982	torsion-balance (period)	6.6726

Possible rates of change

tation might be proportional to the age of the universe; other rates of change over time also have been proposed. The rates of change would be extremely small, one part in 10^{11} per year if the age of the universe is taken to be 10^{11} years; such a rate is entirely beyond experimental capabilities at present. There is, however, the possibility of looking for the effects of any variation upon the orbit of a celestial body, in particular, the Moon. It has been claimed from time to time that such effects may have been detected. As yet, there is no certainty.

Fundamental character of G. The constant of gravitation is plainly a fundamental quantity, since it appears to determine the large-scale structure of the entire universe. Gravity is a fundamental quantity whether it is an essentially geometric parameter as in general relativity or whether it is the strength of a field that is one aspect of a more general field of unified forces. The fact that, so far as is known, gravitation depends on no other physical factors makes it likely that the value of G reflects a basic restriction on the possibilities of physical measurement, just as special relativity is a consequence of the fact that, beyond the shortest distances, it is impossible to make separate measurements of length and time. (A.H.C.)

BIBLIOGRAPHY

General: ISAAC NEWTON, *The Mathematical Principles of Natural Philosophy,* 2 vol. (1729, reissued in 1 vol., 1975; originally published in Latin, 1 vol., 1687), often referred to as the *Principia,* is the origin of all fundamental work on gravity. STEPHEN W. HAWKING and W. ISRAEL (eds.), *Three Hundred Years of Gravitation* (1987), provides many authoritative review articles in commemoration of the tercentenary of the publication of Newton's *Principia.* CHARLES W. MISNER, KIP S. THORNE, and JOHN ARCHIBALD WHEELER, *Gravitation* (1973), is a leading work on gravitational theory. ALAN COOK, *The Motion of the Moon* (1988), discusses theories of the lunar orbit, with a chapter on applications that includes an account of gravitational studies.

Gravity around the Earth, Moon, and planets: ALAN COOK, *Physics of the Earth and Planets* (1973), includes a chapter on methods and results of gravity measurements. WOLFGANG TORGE, *Geodesy: An Introduction* (1980), contains a full and up-to-date chapter on gravity measurements. JAMES A. HAMMOND and JAMES E. FALLER, "Results of Absolute Gravity Determinations at a Number of Sites," *Journal of Geophysical Research,* 76(32):7850–7854 (Nov. 10, 1971), gives an account of absolute measurements by free fall. C. MORELLI *et al., The International Gravity Standardization Net 1971 (I.G.S.N. 71)* (1974), is the result of adjusting a large number of gravity measurements over the Earth. ALAN COOK, *Interiors of the Planets* (1980), summarizes knowledge of the gravity fields of the planets and their interpretation.

Theories of gravitation and astronomical aspects: STEPHEN W. HAWKING, *A Brief History of Time: From the Big Bang to Black Holes* (1988), is a nonmathematical book by an outstanding author; see especially the chapter on black holes. CLIFFORD M. WILL, *Theory and Experiment in Gravitational Physics* (1981, reprinted 1985), is a thorough treatment. See also J.H. TAYLOR and P.M. MCCULLOCH, "Evidence for the Existence of Gravitational Radiation from Measurements of the Binary Pulsar PSR 1913+16," *Annals of the New York Academy of Sciences,* 336:442–446 (1980); and B.W. PETLEY, *The Fundamental Physical Constants and the Frontier of Measurement* (1985, reprinted 1988).

Experiments on gravitation: Experiments on aspects of gravitation are treated in the following journal articles: ALAN COOK, "Experiments on Gravitation," *Reports on Progress in Physics,* 51(5):707–757 (May 1988), a comprehensive review, with many references to early work, to the measurements of G in the table, and to recent studies of the inverse square law; P.G. ROLL, R. KROTKOV, and R.H. DICKE, "The Equivalence of Inertial and Passive Gravitational Mass," *Annals of Physics,* 26(3):442–517 (Feb. 20, 1964), a thorough account of a classical experiment; HENRY CAVENDISH, "Experiments to Determine the Density of the Earth," *Philosophical Transactions of the Royal Society of London,* 88:469–526 (June 21, 1798), the first measurement of G; GABRIEL G. LUTHER and WILLIAM R. TOWLER, "Redetermination of the Newtonian Gravitational Constant G," *Physical Review Letters,* 48(3):121–123 (Jan. 8, 1982), probably the best measurement of G up to that time; F.D. STACEY *et al.,* "Geophysics and the Law of Gravity," *Reviews of Modern Physics,* 59(1):157–174 ((Jan. 1987), a review of work on a possible difference between G as measured in the laboratory and estimated geophysically; and T.C. VAN FLANDERN, "Is the Gravitational Constant Changing?" *The Astrophysical Journal,* 248(2):813–816 (Sept. 1, 1981), a discussion of relevant observations of the Moon.

(A.H.C.)

Greece

The Hellenic Republic (Greek Ellinikí Dimokratía or Ellás) is a medium-sized European nation stretching across the tip of the Balkan Peninsula from the Ionian Sea to the Aegean and including islands in both seas. Its area—50,960 square miles (131,986 square kilometres)—is about that of Czechoslovakia or North Korea, and its neighbours (clockwise from the northwest) are Albania, Yugoslavia, Bulgaria, and Turkey; its insular possessions stretch as far south as the major Mediterranean island of Crete (Kríti). Its capital is Athens (Athínai). The land now known as Greece was inhabited at least 8,000 years ago. The name Greece (Latin Graecia) derives from *Graeci,* originally the Latin name of a Boeotian tribe (*Graioi*) that settled in Italy in the 8th century BC but later applied there to the Hellenic people as a whole. Greeks crossed the Mediterranean to establish significant colonies in Spain, southern France, Italy, Asia Minor, and North Africa. Particularly celebrated were the Greeks (usually called Achaeans by Homer) who sacked Troy in about 1200 BC.

There is about Greece a vitality, all too often undisciplined, that makes many other European countries seem tame, even dull, by comparison. This is evident immediately upon sailing into a Greek port or upon crossing a Greek border: sounds, smells, movements, colours—the very tempo of things—conspire to heighten sensibilities and intensify expectations. But alongside all this is the serene coolness, even aloofness, of what remains in Greece from classical antiquity, the visible monuments of which constantly stand as a challenge to (if not even a rebuke of) contemporary endeavours.

The vitality of Greece and Greeks can be said to stem from the heady mixture over centuries of many peoples and ways of life. The land has long been "at the crossroads"—at that place in the Balkans where "three roads meet," where Europe, Africa, and the East converge, and at that point in time where the ancient, the medieval, and the modern coexist and conflict. Greeks are familiar with the Middle East: their language and food, to say nothing of their religion and history, are marked by exotic and even oppressive elements from Turkey, the Holy Land, Egypt, and beyond. But they are familiar as well with the soberer West, with the Europe to which the young go for training and for work, with the Americas and Australia to which so many have gone for a new home. Greeks, like the Jews, Scots, and Chinese, whom they resemble in so many ways, have long been able to accommodate themselves as merchants in many climes and to many ways of life. But, also like the Jews and perhaps the Chinese, they have preserved in their heart of hearts a vital memory of the homeland to which they yearn to return.

This homeland is as much a state of mind as it is a place to be found on maps. The yearning to "return," then, is almost as strong among those who have never left Greek soil as among those who find themselves abroad. Perhaps it is a yearning to attain that which has never been but which has always been aspired to. It is a yearning evident in the melancholy of Greek music, the nostalgia that can be heard even in the lively tunes and ballads sung on festive occasions. It is a yearning that can be heard as well in interminable conversations, especially those with which Greeks refresh and recreate themselves through the long cool nights that follow blistering summer days. And it is a yearning that can be seen in the faces and deeds of the Greeks, a yearning that makes it impossible for them "to leave well enough alone." A perpetual restlessness, much like that which was said to characterize the political life of ancient Athens, is evident, a restlessness that can continue subterraneanly despite surface conformity to the tyranny of the moment.

Indeed, there have been many tyrannies in Greece, tyrannies that are as much a part of the much-discussed "Greek experience" as (if not even the most frequent result of) their volatile democracies. Perhaps, it might even be said, memories of intermittent tyrannies remind Greeks of the unpredictability of human things, of the disaster that can follow upon prosperity, of the trials that even the most successful encounter from time to time. Life can be expected to be as hard, as unyielding, and as toughening as the soil and the sea from which Greeks have for centuries wrested their livelihood. But it can also be as enriching and as exciting as the landscape and the light for which Greece has always been celebrated and that can be seen, if not at this moment or place, then surely in a little while or down the road a few kilometres. (G.An.)

This article is divided into the following sections:

Physical and human geography 179
- The land 179
  - Relief
  - Climate
  - Drainage
  - Plant and animal life
  - Settlement patterns
- The people 185
  - Linguistic, ethnic, and religious background
  - Demography
- The economy 185
  - Resources
  - Energy
  - Agriculture, forestry, and fishing
  - Industry
  - Finance
  - Trade
  - Transportation
- Government and social conditions 187
  - Government
  - Education
  - Health and welfare
- Cultural life 188

History 189
- Early cultures 189
  - The Bronze Age
  - The Dorian invasions
  - The Archaic period 190
  - The Classical period 190
- Macedonian period after Alexander 191
  - Macedonia and the Greek city-states, 323–224 BC
  - Coming of the Romans
- Postclassical history 192
  - Greece under the Roman Republic
  - Roman imperial rule: Augustus to Diocletian
- Byzantine history 193
  - Early Byzantine period
  - From the Isaurians to the Angeli, 717–1204
  - Late Byzantine period
- Modern history 195
  - Greece under the Turks, 1453–1821
  - Rising of 1770 and beginnings of nationalism
  - The Revolution of 1821 and its results
  - The new monarchy, 1863–1924
  - Balkan wars, 1912–13
  - World War I and Asia Minor
  - The republic, 1924–35
  - Monarchy restored
  - World War II
  - Communist rebellion
  - Reconstruction and recovery
  - Military government, 1967–74
  - Restoration of parliamentary government

Bibliography 204

Physical and human geography

THE LAND

Relief. The Greek landscape is conspicuous not only for its beauty but also for its complexity and its variety. The dominant element—as noted by Strabo, the great geographer of classical antiquity, and confirmed by a glance at the map—is the sea. An ever-present factor, the sea presses deep into the land in a host of arms and inlets, which are often separated by the rocky spines of peninsulas that thrust back into the sea and are continued in the arcs and clusters of beautiful islands across its surface. Only a small, wedge-shaped portion of the interior of modern Greece is more than 50 miles (80 kilometres) from the sea. Mountains are the second major element in Greek topography. They cover about three-quarters of the country's surface, forming a ribbed, interlaced network, trending generally from northwest to southeast and enclosing numerous small basins. The basins—together with narrow valleys, small plains (spreading more extensively about river mouths), and a thin, broken coastal strip—form the third element in the relief, the lowland.

The three elements in the topography

In their combination and interaction, these three elements have been of immense significance in shaping national development: the rich soils of the basins nurtured agriculture and the first stirrings of civilization, but the mountains (while often serving as a barrier against invaders) restricted this social development to small, fiercely independent areas and impeded communications among them. The sea—as both history and the ancient stories attest—made for contact among the separate communities and stimulated contacts farther afield, although it also exposed the coastal regions to external attack. For the geographer these three elements, in local combination, are a convenient basis for a division of the contemporary Greek landscape into six natural regions.

The Pindus Mountains. The core region in Greek topography is unquestionably the rugged Pindus Mountains (Píndhos Óros) area of the northwest interior. Following the general northwest–southeast trend of the mountains of the Balkan Peninsula, the Pindus sweep down in a series of rugged, roughly parallel ridges from the Albanian and Yugoslavian frontier and are structurally a part of the Dinaric mountain system of those countries. This system of fold mountains was created relatively late in geological time, and earthquakes continue to afflict the region as the mountain structures settle down. The highest point in this region is Smólikas Óros (*óros,* plural *óri,* "mountain"), 8,651 feet (2,637 metres) high. The mountain scenery, with jagged granitic peaks, wild gorges, and a succession of magnificent views glimpsed from winding roads, is justly famed.

Robert Harding Picture Library, London



Limestone buttresses overlooking Vikos Gorge in the Pindus Mountains, near Ioánnina, in northwestern Greece.

Macedonia and Thrace. A number of topographic regions surround this mountainous core and are often penetrated by extensions of it. The northernmost division, roughly the regions of Macedonia and Thrace (Thráki), extends in a long, narrow east–west band between the northern Aegean coast and the frontier with Yugoslavia and Bulgaria. It is bounded on the east by the Maritsa River (Évros Potamós; *potamós,* "river"), which marks the Turkish frontier, and consists of hills and forest-clad mountains interpenetrated by valleys, river basins, and alluvial plains. Along the Bulgarian border and beyond it rise the ancient crystalline rock structures of the Rhodope Mountains (Rodhópi Óri), against which the newer mountains of the Pindus were crushed during their formation. In the west the three-pronged peninsula of Chalcidice (Khalkidhikí) forms a distinctive feature as it thrusts out into the Aegean. On the easternmost prong, Áyion Óros (Holy Mountain), is located Mt. Áthos, the famous site of monastic communities.

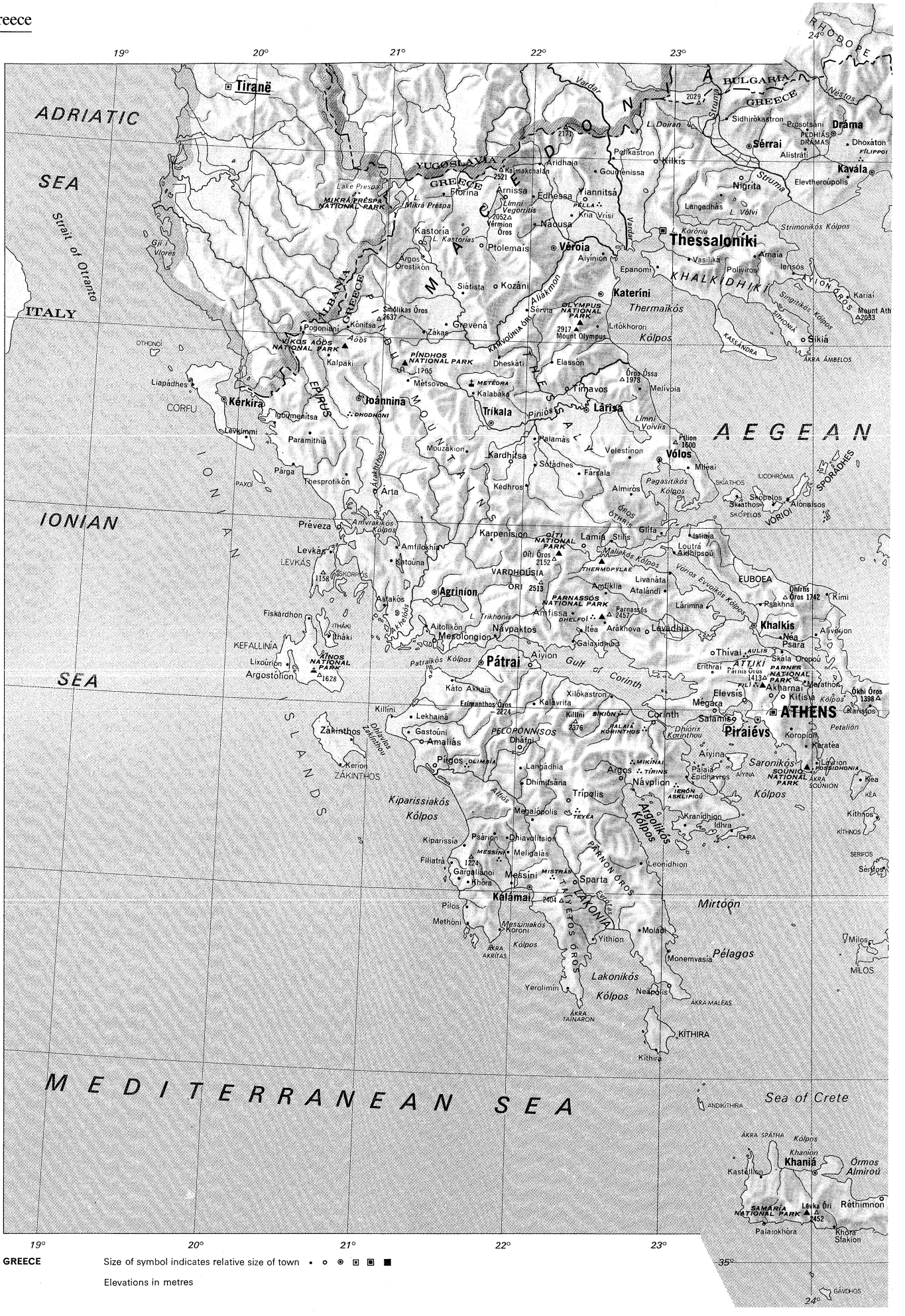
The Chalcidice Peninsula and Mt. Áthos

The peninsula is separated from the rest of the coastal region by a fault line of structural weakness, marked by the lakes Korónia and Vólvi. Just to the west extends a large plain drained by the Vardar (Axiós) and Aliákmon rivers, whose swampy deltas are slowly pushing out into the nearby Thermaïkós Kólpos (*kólpos,* "gulf"). The forested Vérmion Óros and, beyond it, the barren inland basins around lakes Vegorrítis and Kastorías mark the boundary with the Pindus proper. Farther east is a succession of plains, often swampy; that of Sérrai, around the lower Struma (Strimón), and the deltaic plain of the lower Néstos are most significant. Inland basins of structural origin include that of the Pedhiás Dhrámas.

Central Greece. Central Greece lies to the south of Macedonia and Thrace and is lent character by four spurs that thrust out from the main Pindus mass, following the northwest–southeast trend of that region. A number of distinctive basins and plains lie amidst these upland ribs. The northernmost, a rather broken spur called the Kamvoúnia Óri, runs along the coast of the Thermaïkós Kólpos and continues south to form the peninsula bounding one side of the Órmos Vólou (*órmos,* "bay"). One of its peaks is Mt. Olympus (Óros Ólimbos)—the mythical seat of the gods, whose often cloud-topped summit rises to 9,570 feet, the highest point in Greece—and also the equally fine peaks of Óssa and Mt. Pelion (Óros Pílion). The next spur on the west is the Óros Óthris range, which continues across the narrow Dhíavlos Oreón (Oreón Channel) in the northern sector of the long, narrow island of Euboea (Évvoia). Between the two spurs lie the ancient basins (formerly the site of lakes) of Thessaly (Thessalía), Tríkala (Tríkkala), and Lárisa, drained by the Piniós. Just to their south, the basin of Almirós, of similar origin, lies around the Órmos Vólou.

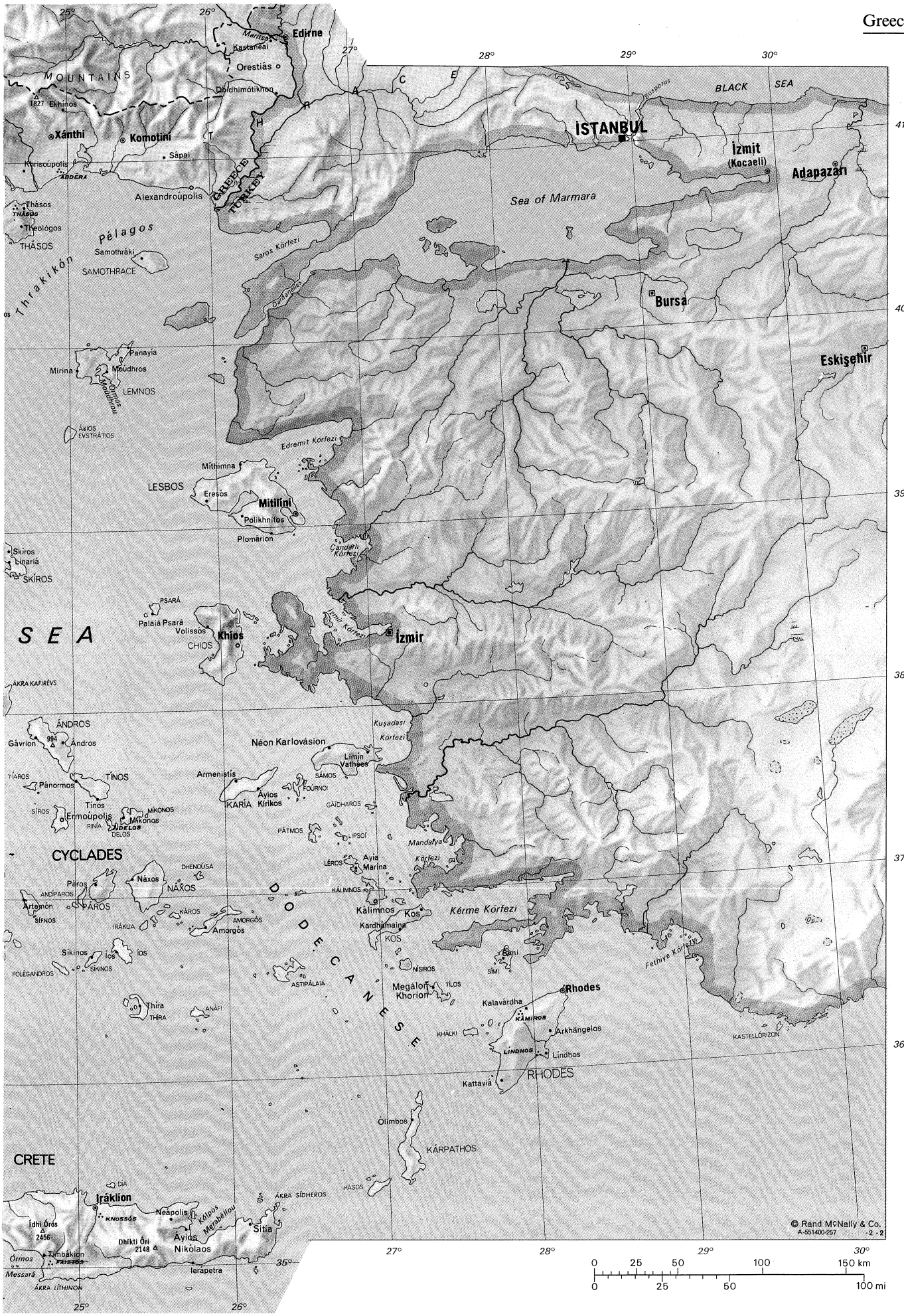
To the southwest, the third spur leaving the Pindus is that of the Oíti, continued in the Ókhi Óros of southern Euboea. Just before the Oíti reaches the sea, near the head of the Maliakós Kólpos, lies the pass of Thermopylae (Thermopílai), scene of the famous battle of antiquity. The last (and perhaps the most important) of the four spurs thrusting down into central Greece is that curving away to the southeast through the twin-peaked mass of Mt. Parnassus (Parnassós). This mountain rises to 8,061 feet and was held to be the home of the Muses. The view from its summit at sunrise, with a broad expanse of the heart of Greece gradually unfolding, is regarded as one of the finest in the world. The range continues as

Mt. Parnassus and its regional context



GREECE Size of symbol indicates relative size of town • ○ ◉ ▣ ▣ ■

Elevations in metres



Edirne
Maritsa
Kastaneai
Orestiás
Dhidhimótikhon
MOUNTAINS
1827 Ekhínos
Xánthi
Komotini
Sápai
Khrisoúpolis
ÁBDERA
Alexandroúpolis
GREECE
TURKEY
Thásos
THÁSOS
Theológos
THÁSOS
Pélagos
Samothráki
SAMOTHRACE
Thrakikón
Saros Körfezi
Dardanelles
Panayía
Mírina
Moúdhros
Ormos Moúdhrou
LEMNOS
ÁYIOS EVSTRÁTIOS
Edremit Körfezi
Mithimna
LESBOS
Eresós
Mitilíni
Polikhnitos
Plomárion
Skíros
Linariá
SKÍROS
PSARÁ
Palaiá Psará
Volissós
Khíos
CHIOS
SEA
ÁKRA KAFIRÉVS
ÁNDROS
Gávrion
994
Ándros
YÍAROS
Pánormos
TÍNOS
Tínos
SÍROS
Ermoúpolis
MÍKONOS
Mikonos
RINÍA
DELOS
DELOS
CYCLADES
Páros
Náxos
DHENOÚSA
NÁXOS
ANDÍPAROS
Artemón
PÁROS
SÍFNOS
KÁROS
IRÁKLIA
AMORGÓS
Amorgós
Síkinos
Íos
Íos
SÍKINOS
FOLÉGANDROS
Thíra
THÍRA
ANÁFI
ASTIPÁLAIA
DODECANESE
Kuşadası Körfezi
Néon Karlovásion
Limín Vathéos
Armenistís
SÁMOS
FOÚRNOI
Áyios Kírikos
IKARÍA
GÁÏDHAROS
PÁTMOS
LIPSOÍ
LÉROS
Ayía Marína
Mandalya Körfezi
KÁLIMNOS
Kálimnos
Kos
Kardhámaina
KOS
Kérme Körfezi
NÍSIROS
Sámi
SÍMI
Megálon Khorion
TÍLOS
Rhodes
Kalavárdha
KÁMIROS
Arkhángelos
KHÁLKI
LÍNDHOS
Líndhos
RHODES
Kattaviá
Fethiye Körfezi
KASTELLÓRIZON
Ólimbos
KÁRPATHOS
KÁSOS
CRETE
DÍA
Iráklion
KNOSSÓS
Neápolis
Kólpos Merabéllou
ÁKRA SÍDHEROS
Sitía
Ídhi Óros
2456
Dhíkti Óri
2148
Áyios Nikólaos
Ormos
Timbákion
FAISTÓS
Messará
Ierápetra
ÁKRA LÍTHINON
BLACK SEA
Bosporus
İSTANBUL
İzmit
(Kocaeli)
Adapazarı
Sea of Marmara
Bursa
Eskişehir
İzmir Körfezi
İzmir
25°
26°
27°
28°
29°
30°
41°
40°
39°
38°
37°
36°
35°
© Rand McNally & Co.
A-551400-257
-2-2
0 25 50 100 150 km
0 25 50 100 mi

MAP INDEX

Cities and towns

Agrinion 38·37n 21·24e
Aitolikón 38·27n 21·22e
Aíyina 37·44n 23·27e
Aiyínion 40·30n 22·33e
Aíyion 38·15n 22·05e
Akharnaí 38·05n 23·44e
Alexandroúpolis 40·50n 25·52e
Alístráti 41·04n 23·57e
Alivérion 38·24n 24·02e
Almirós 39·11n 22·46e
Alónnisos 39·08n 23·50e
Amaliás 37·49n 21·23e
Amfíklia 38·38n 22·35e
Amfilokhía 38·51n 21·10e
Ámfissa 38·31n 22·24e
Amorgós 36·50n 25·54e
Ándros 37·50n 24·57e
Arákhova 38·29n 22·35e
Árgos 37·39n 22·44e
Árgos Orestikón 40·28n 21·16e
Argostólion 38·10n 20·30e
Aridhaía 40·59n 22·03e
Arkhángelos 36·12n 28·08e
Armenistís 37·36n 26·08e
Arnaía 40·29n 23·35e
Árnissa 40·48n 21·50e
Árta 39·09n 20·59e
Artemón 36·59n 24·43e
Astakós 38·32n 21·05e
Atalándi 38·39n 23·00e
Athens (Athínai) 37·58n 23·43e
Ayía Marína 37·09n 26·52e
Áyios Kírikos 37·37n 26·14e
Áyios Nikólaos 35·11n 25·42e
Corinth (Kórinthos) 37·56n 22·56e
Dháfni 37·48n 22·01e
Dheskáti 39·55n 21·49e
Dhiavolítsion 37·18n 21·58e
Dhidhimótikhon 41·21n 26·30e
Dhimitsána 37·37n 22·03e
Dhoxáton 41·05n 24·14e
Dráma 41·09n 24·08e
Édhessa 40·48n 22·03e
Ekhínos 41·17n 24·59e
Elassón 39·54n 22·11e
Elevsís 38·02n 23·32e
Elevtheroúpolis 40·55n 24·16e
Epanomí 40·26n 22·56e
Erétria, see Néa Psará
Eresós 39·10n 25·56e
Erithraí 38·13n 23·19e
Ermoúpolis 37·26n 24·56e
Fársala 39·18n 22·23e
Filiatrá 37·10n 21·35e
Fiskárdhon 38·27n 20·35e
Flórina 40·47n 21·24e
Galaxídhion 38·22n 22·23e
Gargaliánoi 37·04n 21·39e
Gastoúni 37·51n 21·16e
Gávrion 37·52n 24·46e
Glífa 38·57n 22·58e
Gouménissa 40·57n 22·27e
Grevená 40·05n 21·25e
Ídhra 37·20n 23·29e
Ierápetra 35·00n 25·45e
Ierisós 40·24n 23·52e
Igoumenítsa 39·30n 20·16e
Ioánnina 39·40n 20·50e
Ios 36·44n 25·17e
Iráklion 35·20n 25·08e
Istiáía 38·57n 23·09e
Itéa 38·26n 22·24e
Ithráki 38·23n 20·42e
Kalabáka 39·42n 21·38e
Kalámai 37·04n 22·07e
Kalavárdha 36·20n 27·57e
Kalávrita 38·01n 22·06e
Kálimnos 36·57n 26·59e
Kalpáki 39·53n 20·38e
Kardhámaina 36·47n 27·09e
Kardhítsa 39·21n 21·55e
Kariaí 40·16n 24·15e
Káristos 38·00n 24·24e
Karpenísion 38·55n 21·47e
Kastanéai 41·38n 26·28e
Kastéllion 35·30n 23·38e
Kastoría 40·31n 21·15e
Kateríni 40·16n 22·30e
Káto Akhaïa 38·09n 21·32e
Katoúna 38·47n 21·07e
Kattaviá 35·57n 27·46e
Kavála 40·56n 24·25e
Kéa 37·38n 24·21e
Kédhros 39·13n 22·03e
Keratéa 37·48n 23·59e
Kerío 37·40n 20·48e
Kérkira 39·36n 19·56e
Khalkís 38·28n 23·36e
Khaniá 35·31n 24·02e
Khíos 38·22n 26·08e
Khóra 37·04n 21·43e
Khóra Sfakíon 35·12n 24·09e
Khrisoúpolis 40·58n 24·42e
Kifisiá 38·04n 23·48e
Kilkís 41·00n 22·53e
Killíni 37·55n 21·09e
Kímí 38·37n 24·06e
Kiparissía 37·14n 21·40e
Kíthira 36·09n 22·58e
Kíthnos 37·26n 24·26e
Komotiní 41·08n 25·25e
Kónitsa 40·02n 20·45e
Koróni 36·48n 21·56e
Koropíon 37·54n 23·53e
Kos 36·53n 27·18e
Kozáni 40·18n 21·47e
Kranídhion 37·22n 23·10e
Kría Vrísi 40·41n 22·18e
Lamía 38·54n 22·26e
Langadhás 40·45n 23·04e
Langádhia 37·41n 22·02e
Lárimna 38·34n 23·18e
Lárisa 39·38n 22·25e
Lávrion 37·44n 24·04e
Lekhainá 37·56n 21·17e
Leonídhion 37·10n 22·52e
Levádhia 38·25n 22·54e
Levkás 38·50n 20·41e
Levkímmi 39·25n 20·04e
Liapádhes 39·40n 19·44e
Limín Vathéos 37·45n 27·00e
Linariá 38·50n 24·32e
Líndhos 36·06n 28·04e
Litókhoron 40·06n 22·30e
Livanátai 38·42n 23·03e
Lixoúrion 38·12n 20·26e
Loutrá Aidhipsoú 38·51n 23·02e
Marathón 38·10n 23·58e
Megálon Khoríon 36·27n 27·21e
Megalópolis 37·24n 22·08e
Mégara 38·01n 23·21e
Meligalás 37·13n 21·59e
Melívoia 39·45n 22·48e
Mesolóngion 38·21n 21·26e
Messíni 37·04n 22·00e
Methóni 36·50n 21·43e
Métsovon 39·46n 21·11e
Míkonos 37·26n 25·20e
Miléai 39·20n 23·09e
Mílos 36·45n 24·27e
Mírina (Kástron) 39·52n 25·04e
Mítimna 39·22n 26·10e
Mitilíni 39·06n 26·32e
Moláoi 36·48n 22·52e
Monemvasía 36·41n 23·03e
Moúdhros 39·52n 25·16e
Mouzákion 39·26n 21·40e
Náousa 40·37n 22·05e
Návpaktos 38·23n 21·50e
Návplion 37·34n 22·48e
Náxos 37·06n 25·23e
Neápolis 35·15n 25·37e
Neápolis 36·30n 23·04e
Néa Psará 38·24n 23·48e
Néon Karlovásion 37·48n 26·44e
Nigríta 40·55n 23·30e
Ólimbos 35·44n 27·11e
Orestiás 41·30n 26·31e
Palaiá Epídhavros 37·38n 23·09e
Palaiá Psará 38·32n 25·34e
Palaiokhóra 35·14n 23·41e
Palamás 39·28n 22·05e
Panayía 39·59n 25·24e
Pánormos 37·38n 25·02e
Paramithiá 39·28n 20·30e
Párga 39·17n 20·23e
Páros 37·04n 25·08e
Pátrai 38·15n 21·44e
Pílos 36·55n 21·43e
Piraiévs (Piraeus) 37·57n 23·38e
Pírgos 37·41n 21·28e
Plomárion 38·59n 26·22e
Pogonianí 40·00n 20·25e
Polikástron 41·00n 22·34e
Polikhnítos 39·05n 26·11e
Políyiros 40·23n 23·27e
Préveza 38·57n 20·44e
Prosotsáni 41·10n 23·59e
Psakhná 38·35n 23·38e
Psárion 37·20n 21·51e
Ptolemaís 40·31n 21·41e
Réthimnon 35·22n 24·29e
Rhodes (Ródhos) 36·26n 28·13e
Salamís 37·59n 23·28e
Salonika, see Thessaloníki
Samothráki 40·28n 25·31e
Sápai 41·02n 25·41e
Sérifos 37·09n 24·31e
Sérrai 41·05n 23·32e
Sérvia 40·11n 22·00e
Siátista 40·16n 21·33e
Sidhirókastron 41·14n 23·22e
Sikiá 40·02n 23·56e
Síkinos 36·42n 25·07e
Símí 36·36n 27·50e
Sitía 35·12n 26·07e
Skála Oropoú 38·20n 23·46e
Skíathos 39·10n 23·29e
Skíros 38·53n 24·33e
Skópelos 39·07n 23·43e
Sofádhes 39·20n 22·06e
Sparta (Spárti) 37·05n 22·27e
Stilís 38·55n 22·36e
Thásos 40·47n 24·42e
Theológos 40·39n 24·41e
Thesprotikón 39·15n 20·47e
Thessaloníki (Salonika) 40·38n 22·56e
Thíra 36·25n 25·26e
Thívai 38·21n 23·19e
Timbákion 35·04n 24·46e
Tínos 37·32n 25·10e
Tírnavos 39·45n 22·17e
Tríkala 39·34n 21·46e
Trípolis 37·31n 22·21e
Vasiliká 40·28n 23·08e
Velestínon 39·23n 22·45e
Véroia 40·31n 22·12e
Volissós 38·29n 25·58e
Vólos 39·21n 22·56e
Xánthi 41·08n 24·53e
Xilókastron 38·05n 22·38e
Yerolimín 36·28n 22·24e
Yiannitsá 40·48n 22·25e
Yíthion 36·45n 22·34e
Zákas 40·02n 21·16e
Zákinthos 37·47n 20·53e

Physical features and points of interest

Abdera, *ruins* 40·59n 24·58e
Aegean Sea 38·30n 25·00e
Aínos National Park 38·09n 20·46e
Aíyina, *island* 37·46n 23·26e
Akhelóös, *river* 38·36n 21·14e
Akrítas, Ákra, *cape* 36·43n 21·54e
Alfiós, *river* 37·37n 21·27e
Aliákmon, *river* 40·30n 22·36e
Almiroú, Órmos, *bay* 35·23n 24·20e
Ámbelos, Ákra, *point* 39·56n 23·55e
Amorgós, *island* 36·50n 25·59e
Amvrakikós Kólpos, *bay* 39·00n 21·00e
Anáfi, *island* 36·21n 25·50e
Andikíthira, *island* 35·52n 23·18e
Andíparos, *island* 37·00n 25·03e
Ándros, *island* 37·50n 24·50e
Aöös, *river* 40·37n 19·20e
Árakhthos, *river* 39·01n 21·03e
Argolikós Kólpos, *gulf* 37·20n 22·55e
Astipálaia, *island* 36·35n 26·25e
Athos, Mount (Áthos), *mountain* 40·09n 24·19e
Attikí, *historic region* 38·00n 23·30e
Aulís, *ruins* 38·27n 23·35e
Áyion Óros (Áthos), *peninsula* 40·15n 24·15e
Áyios Evstrátios, *island* 39·30n 25·04e
Chios (Khíos), *island* 38·22n 26·00e
Corfu (Kérkira), *island* 39·40n 19·45e
Corinth, Gulf of (Korinthiakós Kólpos) 38·12n 22·30e
Cos, see Kos
Crete (Kríti), *island* 35·15n 24·45e
Crete, Sea of (Kritikón Pélagos) 36·00n 25·00e
Cyclades (Kikládhes), *islands* 37·00n 25·10e
Cythera, see Kíthira
Delos (Dhílos), island 37·22n 25·16e
Delos, *ruins* 37·22n 25·16e
Delphi, see Dhírfis Óros
Dhelfoí, *ruins* 38·30n 22·29e
Dhenoúsa, *island* 37·06n 25·50e
Dhíkti Óri, *mountains* 35·08n 25·22e
Dhírfis Óros (Delphi), *mountain* 38·38n 23·49e
Dhodhóni, *ruins* 39·34n 20·47e
Día, *island* 35·28n 25·14e
Dodecanese (Dhodhekánisos, Sporádhes), *islands* 36·00n 27·00e
Doiran, Lake (Doïráni) 41·13n 22·44e
Drámas, Pedhiás, *plain* 41·04n 24·00e
Epirus (Ípiros), *historic region* 39·40n 20·40e
Erímanthos Óros, *mountain* 37·59n 21·51e
Euboea (Évvoia), *island* 38·34n 23·30e
Evrótas, *river* 36·48n 22·40e
Faistós, *ruins* 35·01n 24·48e
Filí, *ruins* 38·10n 23·40e
Filippoi, *ruins* 41·00n 24·16e
Folégrandros, *island* 36·36n 24·56e
Foúrnoi, *island* 37·34n 26·30e
Gáïdharos, *island* 37·28n 27·00e
Gávdhos, *island* 34·50n 24·06e
Hydra, see Ídhra
Ídhi Óros, *mountain* 35·18n 24·43e
Ídhra (Hydra), *island* 37·20n 23·32e
Ierón Asklipioú, *ruins* 37·37n 23·02e
Ikaría, *island* 37·35n 26·10e
Iliodhrómia, *island* 39·14n 23·54e
Ionian Islands (Iónioi Nísoi) 38·30n 20·30e
Ionian Sea 38·30n 19·00e
Íos, *island* 36·42n 25·24e
Iráklia, *island* 36·50n 25·26e
Ithráki, *island* 38·24n 20·42e
Kafirévs, Ákra, *cape* 38·09n 24·36e
Kaïmakchalán, *mountain* 40·58n 21·48e
Kálimnos, *island* 37·00n 27·00e
Kámiros, *ruins* 36·19n 27·57e
Kamvoúnia Óri, *mountains* 40·00n 21·52e
Káros, *island* 36·52n 25·40e
Kárpathos, *island* 35·40n 27·10e
Kásos, *island* 35·22n 26·56e
Kassándra, *peninsula* 40·00n 23·20e
Kastellórizon, *island* 36·08n 29·34e
Kastorías, Límni, *lake* 40·30n 21·17e
Kéa, *island* 37·34n 24·22e
Kefallinía, *island* 38·15n 20·35e
Khálki, *island* 36·17n 27·35e
Khalkidhikí, *historic region* 40·25n 23·27e
Khaníon, Kólpos, *bay* 35·34n 23·48e
Killíni, *mountain* 37·57n 22·23e
Kiparissiakós Kólpos, *bay* 37·30n 21·25e
Kíthira (Cythera), *island* 36·15n 23·00e
Kíthnos, *island* 37·25n 24·28e
Knossós, *ruins* 35·20n 25·10e
Korínthou, Dhiórix, *canal* 37·57n 22·56e
Korónia, Límni, *lake* 40·41n 23·05e
Kos (Cos), *island* 36·50n 27·10e
Lakonía, *historic region* 37·00n 22·30e
Lakonikós Kólpos, *bay* 36·35n 22·40e
Lamia, Gulf of, see Maliakós Kólpos
Lemnos (Límnos), *island* 39·55n 25·15e
Léros, *island* 37·08n 26·52e
Lesbos (Lésvos), *island* 39·10n 26·32e
Lévka Óri, *mountains* 35·18n 24·01e
Levkás, *island* 38·43n 20·38e
Límnos, see Lemnos
Líndhos, *ruins* 36·06n 28·05e
Lipsoí, *island* 37·20n 26·45e
Líthinon, Ákra, *cape* 34·55n 24·44e
Macedonia, *historic region* 41·00n 23·00e
Maléas, Ákra, *cape* 36·26n 23·12e
Maliakós Kólpos (Gulf of Lamia) 38·52n 22·38e
Maritsa (Évros), *river* 40·52n 26·12e
Mediterranean Sea 36·00n 21·00e
Merabéllou, Kólpos, *bay* 35·14n 25·47e
Messará, Órmos, *bay* 34·58n 24·36e

MAP INDEX (continued)

Messíni, *ruins* . . . 37·11n 21·57e
Messiniakós Kólpos (Gulf of Messinia), *bay* . . 36·45n 22·10e
Metéora, *monastery* . . . . . 39·46n 21·36e
Mikínai, *ruins* . . . . 37·44n 22·45e
Míkonos, *island* . . 37·29n 25·25e
Mikrá Préspa, Límni, *lake* . . . . . 40·46n 21·04e
Mikrá Préspa National Park . . . 40·45n 21·06e
Mílos, *island* . . . . 36·41n 24·25e
Mirtóön Pélagos, *sea* . . . . . 37·00n 24·00e
Mistrás, *ruins* . . . . 37·04n 22·21e
Morea, see Pelopónnisos
Moúdhrou, Órmos, *bay* . . . . 39·49n 25·14e
Náxos (Naxía), *island* . . . . . 37·02n 25·35e
Néstos, *river* . . . . 40·51n 24·44e
Nísiros, *island* . . . 36·35n 27·10e
Oíti National Park. 38·50n 22·18e
Oíti Óros, *mountain* . . . . . 38·49n 22·17e
Ókhi Óros, *mountain* . . . . . 38·05n 24·25e
Olimbía, *ruins* . . . 37·38n 21·41e
Olympus, Mount (Ólimbos), *mountain* . . . . . 40·05n 22·21e
Olympus National Park. . . . . 40·04n 22·24e
Óssa, Óros, *mountain* . . . . . 39·49n 22·42e
Othonoí, *island* . . 39·50n 19·26e
Óthris, Óros, *mountains* . . . . . 39·02n 22·37e
Pagasitikós Kólpos (Gulf of Volos), *bay* . . . . 39·15n 22·51e
Palaiá Kórinthos, *ruins* . . . . . 37·54n 22·56e
Parnassós, *mountain* . . . . . 38·32n 22·35e
Parnassós National Park . . . 38·33n 22·33e
Parnes National Park. . . . . 38·10n 23·43e
Párnis Óros, *mountain* . . . . . 38·11n 23·42e
Párnon Óros, *mountains* . . . . . 37·18n 22·35e
Páros, *island* . . . . 37·08n 25·12e
Pátmos, *island* . . . 37·20n 26·33e
Patraïkós Kólpos (Gulf of Patras), *bay* . . . . . 38·15n 21·30e
Paxoí, *island* . . . . 39·12n 20·12e
Pélla, *ruins* . . . . . 40·45n 22·33e
Pelopónnisos (Morea), *historic region* . . . . . 37·30n 22·00e
Petalión, Kólpos, *bay* . . . . . 37·59n 24·02e
Pílion, *mountain* . . 39·28n 23·02e
Píndhos National Park. . . . . 39·54n 21·08e
Pindus Mountains (Píndhos Óros) . 39·49n 21·14e
Piniós, *river* . . . . . 39·54n 22·45e
Possidhonia, *ruins* . . . . . 37·40n 24·00e
Prespa, Lake . . . . 40·50n 21·02e
Psará, *island* . . . . 38·35n 25·37e
Rhodes (Ródhos), *island* . . . . . 36·10n 28·00e
Rhodope, *mountains* . . . . . 41·30n 24·30e
Saloníki, Gulf of, see Thermaïkós Kólpos
Samaría National Park. . . . . 35·17n 23·51e
Sámos, *island* . . . 37·48n 26·44e
Samothrace (Samothráki), *island*. . . . . 40·30n 25·32e
Saronikós Kólpos, *gulf* . . . . . 37·45n 23·30e
Sérifos, *island* . . . 37·11n 24·31e
Sídheros, Ákra, *cape* . . . . . 35·19n 26·19e
Sífnos, *island* . . . . 36·59n 24·40e
Sikión, *ruins* . . . . . 37·59n 22·44e
Síkinos, *island* . . . 36·39n 25·06e
Símí, *island* . . . . . 36·35n 27·52e
Singitikós Kólpos, *bay* . . . . 40·12n 23·56e
Síros, *island* . . . . 37·26n 24·54e
Sithoniá, *peninsula* . . . . . 40·05n 23·55e
Skíathos, *island* . . 39·12n 23·28e
Skíros, *island* . . . . 38·53n 24·32e
Skópelos, *island* . 39·10n 23·40e
Skorpiós, *island* . . 38·41n 20·45e
Smólikas Óros, *mountain* . . . . . 40·06n 20·52e
Soúnio National Park. . . . . 37·43n 23·49e
Soúnion, Ákra, *cape* . . . . . 37·39n 24·02e
Spátha, Ákra *cape* . . . . . 35·42n 23·44e
Strimonikós Kólpos, *bay* . . . . 40·40n 23·50e
Struma (Strimón), *river*. . . . . 40·47n 23·51e
Taínaron, Ákra, *cape* . . . . . 36·22n 22·30e
Taíyetos Óros, *mountains* . . . . . 37·06n 22·18e
Teyéa, *ruins* . . . . . 37·29n 22·24e
Thásos, *island* . . . 40·41n 24·47e
Thásos, *ruins* . . . . 40·46n 24·33e
Thermaïkós Kólpos (Gulf of Saloníki), *bay* . . . 40·23n 22·47e
Thermopylae (Thermopílai), *battlefield*. . . . . 38·48n 22·33e
Thessaly (Thessalía), *historic region* . . 39·30n 22·00e
Thíra, *island*. . . . . 36·24n 25·29e
Thrace (Thráki), *historic region* . . 41·20n 26·45e
Thrakikón Pélagos, *sea* . . . 40·15n 24·28e
Tílos, *island* . . . . . 36·25n 27·25e
Tínos, *island* . . . . 37·38n 25·10e
Tírins, *ruins* . . . . . 37·39n 22·46e
Trikhonís, Límni, *lake* . . . . . 38·34n 21·28e
Vardar (Axiós), *river* . . . 40·35n 22·50e
Vardhoúsia Óri, *mountains* . . . . . 38·44n 22·07e
Vegorrítis, Límni, *lake* . . . . . 40·41n 21·44e
Vérmion Óros, *mountain* . . . . . 40·39n 21·53e
Víkos Aóös National Park . . . 39·58n 20·45e
Voïvïïs, Límni, *lake* . . . . . 39·32n 22·45e
Volos, Gulf of, see Pagasitikós Kólpos
Vólvi, Límni, *lake* . 40·41n 23·23e
Vórioi Sporádhes, *islands*. . . . . 39·15n 23·55e
Vórios Evvoïkós Kólpos, *bay* . . . . 38·40n 23·15e
Yíaros, *island*. . . . 37·38n 24·44e
Zákinthos (Zante), *island* . . . . . 37·52n 20·44e
Zakínthou, Dhíavlos, *strait* . . 37·50n 21·00e

the backbone of the peninsula lying between the Vórios Evvoïkós Kólpos and the Gulf of Corinth (Korinthiakós Kólpos), and it reaches as far as Párnis Óros, just to the north of Athens. To its north lie the plains of Phocis (Fokís) and Boeotia (Voiotía), and around its southern tip lie the depressions of Attica (Attikí), hotter and more arid but with a strategic importance that helps to explain the rise of Athens.

The Peloponnese. The entire southern portion of mainland Greece forms a peninsula lying to the south of the Gulf of Corinth. Technically, this region, the Peloponnese, or Pelopónnisos, also known as the Morea, is now an island, for the 3.9-mile Corinth Canal cuts across the narrow neck of land formerly separating the Gulf of Corinth from that of Aegina (Aíyina). The Peloponnese consists of an oval-shaped mass with mountains rising to 7,800 feet and four peninsular prongs pushing out southward toward Crete. The limestone mass of the plateau of Arcadia (Arkadhía), where streams disappear underground in the soluble rocks, forms the heart of this mass, with the barren land of Taïyetos Óros (rising 7,897 feet) extending southward to form the backbone of one of the subsidiary peninsulas. This mountainous core is surrounded by a thin fringe of fertile coastal plain in the north and west and by the larger alluvial depressions of Laconia (Lakonía), Messenia (Messinía), and Árgos, which head the inlets between the peninsulas to the south. The coast is indented and offers some fine harbours, and the whole region is noted for its wild beauty.

The western uplands and islands. The western side of the Greek mainland north of the Gulf of Corinth to the Albanian frontier and the offshore islands (the Ionian Islands; Iónioi Nísoi) possess their own distinctive topography and regional character. This effect has been enhanced by the fact that the inland mountainous barrier of the Pindus and the ameliorating climatic influences from the west have led to historic isolation from the rest of Greece. Fertile basins are not as well developed as in some other parts of Greece because they are constricted by the parallel ranges of coastal mountains, but the mountain regions themselves, being adequately supplied with rainfall, are not so barren as elsewhere. Corfu (Kérkira), the island lying opposite the Albanian frontier, is the northernmost of the seven major Ionian Islands. It is fertile and well watered. The other islands, Paxos (Paxoí), Leucas (Levkás), Skorpiós, Ithaca (Itháki), Cephalonia (Kefallinía), and Zacynthus (Zákinthos), lie farther south. Lack of rainfall accentuates their gaunt, broken limestone relief, although Leucas and Zacynthus have sheltered eastern plains.

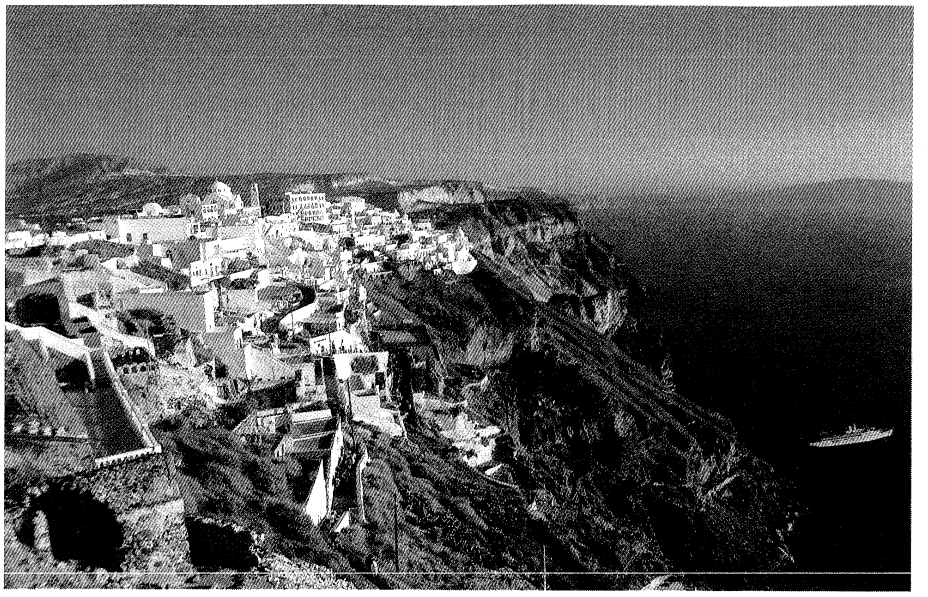
The Aegean Islands. Just as ridgelike extensions of the Pindus interpenetrate the basins and coastal plains of mainland Greece, island groups, which are often further extensions of the same mountain chains, form distinct regional clusters in the Aegean Sea. In the northeast, the island region forms the most extensive—and, visually, perhaps the most attractive—of the physiographic divisions of Greece. In the north, off Thrace (Thráki), lie Thásos (an oval block of ancient mineral rocks similar in composition to neighbouring blocks on the mainland) and harbourless Samothrace (Samothráki), an island of volcanic origin. Lemnos (Límnos), situated midway between Asia Minor and Áyion Óros, is almost cut in two by the northern Órmos Pourniá and the deep southern harbour afforded by the Kólpos Moúdhrou.

Regional clusters of islands

To the southeast, the rocky but sheltered islands of Lesbos (Lésvos), Chios (Khíos), and Sámos lie close to the Turkish coast and are extensions of peninsulas on the coast of Asia Minor. Across the central Aegean, near northern Euboea, lie the Northern Sporades (Vórioi Sporádhes), or "scattered" islands; their crystalline rocks are similar to those of the Greek mainland. Farther south, in the heart of the Aegean, lie the Cyclades (Kikládhes), "islands in a circle." These roughly centre on Delos (Dhílos) and represent the tips of drowned mountain ridges continuing the structural trends of Euboea and the region around Athens.

Between the Cyclades and the Turkish coast, the Dodecanese (Dhodhekánisos) group, with Rhodes (Ródhos) the largest of a dozen major islands, have a varied geological structure ranging from the gray limestones of Kálimnos, Símí, and Khálki to the complete ancient volcanic cone that forms Nísiros. (Ed.)

Finally, the long, narrow shape of Crete stands at the entrance of the Aegean in the extreme southern portion of Greek territory. Crete is dominated by harsh mountains rising out of the sea, stark evidence of its geological origins, for the island is a remnant of a block thrown up in Tertiary times (65,000,000 to 2,500,000 years ago). Much of Crete's 650 miles of rocky coastline slopes down from the major mountains of Crete's east–west axis, a spine that breaks naturally into four main groups: the westernmost Lévka Óri (White Mountains); the central Ídhi (or Psilorítis) Mountains, with Crete's highest point, the summit of Mt. Ídhi, Stavros, 8,058 feet (2,456 metres) high; the



Sunlight on the whitewashed walls of Thíra, chief town of Thíra, the southernmost island of the Cyclades group of the Aegean Islands.
Ivan Massar—Black Star

east central Dhíkti (or Lasíthi) Mountains; and the far eastern Thriftí Mountains. Another range, the Asterousia (or Kófinos) Mountains, runs along the south central coast between the Mesará Plain and the Libyan Sea (Libikon Pélagos). The more gradual slope of the northern coast provides several natural harbours as well as coastal plains, where major cities have grown up: Khaniá, Réthimnon, and Iráklion (the Candia of history). Cretans have lived mostly on the edges of these plains to avoid seasonal flooding, while taking maximum advantage of the arable land.

Crete also is marked with many fissures and ravines, such as the gorge of Samaria, which extends about 11 miles inland from the southwestern coast. There are also caves, the source of many mythological and historical episodes. One freshwater lake, Kournás, lies west of Réthimnon. The island experiences occasional earth tremors, and two serious earthquakes have occurred in the last 100 years.

(J.S.Bo.)

Climate. The basically Mediterranean climate of Greece is subject to a number of regional and even local variations occasioned by the country's physical diversity. In winter the belt of low-pressure disturbances moving in from the North Atlantic shifts southward, bringing with it warm, moist, westerly winds. Squalls and spells of rain ruffle the Aegean, but sunshine often breaks through the clouds. As the low-pressure areas enter the Aegean region, they may draw in cold air from those eastern regions of the Balkans that, sheltered by the Dinaric mountain system from western influences, are open to climatic extremes emanating from the heart of Eurasia. This icy wind is known as the boreas. Partly as a result, Thessaloníki (Salonika) has an average January temperature of 43° F (6° C), while Athens has 50° F (10° C) and Iráklion 54° F (12° C). *Shilok,* or warm winds, are similarly drawn in from the south. The western influences bring plentiful rain to the Ionian coast and the mountains behind it; winter rain also starts early, and snow lingers into spring. At Corfu, January temperatures average 50° F, and the island's average annual rainfall is 52 inches (1,320 millimetres), compared with the total on Crete of 25 inches and the total on Athens of 16 inches. On Crete, snow is almost permanent on the highest peaks.

In summer, when the low-pressure belt swings away again, the climate is hot and dry almost everywhere, with the average July sea-level temperature approaching 80° F (27° C), although heat waves can push the temperature up over the 100° F mark for a day or so. Topography is again a modifying factor: the interior northern mountains continue to experience some rainfall, while all along the winding coast the afternoon heat is eased slightly by sea breezes. In other regions, such as Crete, the hot, dry summers are accentuated by the parching *neltemi,* or etesian winds, which become drier and drier as they are drawn southward.

The quality of Greek light

In all seasons—perhaps especially in summer—the quality of the light is one of Greece's greatest treasures. Although the larger cities have not escaped the pernicious effects of industrial and vehicular pollution, the Greek atmosphere is generally pure and clear. The interplay of light and varied landscape is remarkable. The harsh white limestone crags of the islands, contrasting with the deep blue of the Aegean waters and an equally powerful sky; the dusty green olive groves, the burnt-orange tiled roofs, and dazzling whitewashed walls of coastal communities; and the ever-present weathered stones of the country's great number of ancient monuments all add their own tones.

Drainage. The drainage pattern of Greece is significantly influenced by the porosity and solubility of the rocks of the limestone regions; hence, seasonal downpours are often immediately lost through seepage and runoff. Much rainfall is also lost in rugged terrain of the geologically young northern mountains, where there is a torturous network of rushing mountain streams, often falling into narrow, spectacular gorges. Finally, the irregular, deeply penetrating coastline makes for short river courses. The overall effect is to produce short rivers with an erratic seasonal flow, virtually useless for navigation and limited for irrigation purposes. The Vardar, the Struma, and the Néstos, which cross Macedonia and Thrace to enter the northern Aegean, are the major rivers—but only because they drain large regions beyond the Greek frontier. A host of small and medium-sized rivers drain the rest of the country: the Aliákmon, the Piniós (running east across the main peninsula), and the Evrótas of the Peloponnese are noteworthy.

Plant and animal life. Like other Balkan countries, Greece is open to influences from several major biogeographic zones, with the major Mediterranean influences supplemented by plants and animals stemming from the central European interior. Hence, local topographic and climatic conditions also occasion great variety. On the mountain flanks, and in the north generally, the central European types of vegetation prevail. In central and southern regions and in narrow belts along the valleys of the mountains, about half the land is under scrub of various kinds; and maquis, the classic Mediterranean scrub complex—with oleander, bay, evergreen oak, olive, and juniper—is particularly well developed in the Peloponnese. Evergreen trees and shrubs and herbaceous plants are found in the lowlands, with the flowers offering bril-

liant patterns in springtime. Pines, planes, and poplars line the rivers, the higher slopes, and the coastal plains. Oak, chestnut, and other deciduous trees are found in the north, giving way at higher altitudes to coniferous forests dominated by the Grecian fir, in which clearings are carpeted in spring and summer with irises, crocuses, and tulips. Forests and scrub are found at the highest levels: the black-pine forests covering Mt. Olympus are especially noteworthy.

The forested zones, especially in the north, harbour such European animals as wildcat, martin, brown bear, roe deer, and, more rarely, wolf, wild boar, and lynx. Animals of the Mediterranean regions include jackals, wild goats, and porcupines, all adapted to lack of moisture and to the heat. Birds include pelicans, storks, and herons, while many varieties winter in Greece from farther north. Reptile and fish life is rich and varied.

Settlement patterns. The large number of monuments dotted across Greece are testimony of the antiquity of people's attempts to wrest a living from this sometimes harsh environment. Contemporary patterns of settlement, especially in the rural areas, bear the marks of long centuries of development.

Village life remains a powerful influence. It nevertheless has a cosmopolitan nature—seen routinely in the village-square discussions—while the modern cities have something of the village in their character. The rural communities themselves range from the little communities of the northern mountain interior, reminiscent of central Europe, to the sun-beaten villages of Crete, which are almost African in appearance. Although rural settlement patterns reflect the vagaries of physical geography and the changing currents of history in a particular region, the tiled roofs, the low, whitewashed walls, the long, narrow windows, the central church, and the traces of fortifications are frequent features of village communities. Between villages, the ancient stone walls and winding roads add their own character to the landscape.

Yet town and city life is becoming increasingly important, and about two-thirds of the people live in urban areas. The metropolitan sprawl centred on Athens is one of the great urban complexes of the Mediterranean, complete with industrial and port facilities. The port of Thessaloníki is the second major urban centre. Iráklion is the cosmopolitan major urban area of Crete. Urbanization and the modernization of the economy—especially improvements aimed at the influx of tourists—are also leaving their mark on the traditional Greek landscape.

THE PEOPLE

Linguistic, ethnic, and religious background. Despite the great variety of influences that have shaped modern Greece, and the marked differences among the many isolated regions, a sense of community binds the Greeks together, especially in a national emergency, such as World War II. A common religion, a great heritage, and a common vernacular (variations in regional dialects notwithstanding) tend to make all Greeks feel that they are somehow one people equal to each other in important respects. Combined with this sense of equality is the interest in, and appetite for, political discussion.

The sense of community

All but a very small part of the people belong to the Church of Greece (Greek Orthodox Church). This body appoints its own ecclesiastical hierarchy and is headed by a synod of 12 metropolitans under the presidency of the archbishop of Athens. The Greek Church has links in dogma with the other Orthodox churches. Virtually all Cretans belong to a special branch of the Church of Greece, headed by the archbishop of Crete and directly responsible to the patriarchate of Constantinople.

The Muslim minority, which constitutes most of the non-Orthodox group, is mainly Turkish and is concentrated in western Thrace and the Dodecanese. Roman and Greek Catholics, concentrated in Athens and the western islands formerly under Italian sway, account for the rest, except for a few thousand adherents of Protestant churches, the Anglican Church of Greece, and Judaism, the last having been much reduced in numbers by the Nazi genocide of World War II.



Greek Orthodox priests from the monastery of St. John the Theologian celebrating an outdoor Easter service on Pátmos, island of the Dodecanese group, Aegean Sea.

© John Sims—CLICK/Chicago

In terms of ethnic composition, Greeks again make up all but a small part of the total, the remainder being composed of Macedonians, Turks, Albanians, Bulgarians, Armenians, and Gypsies. Except in Cyprus, southern Albania, and Turkey, there are no major enclaves of Greeks in nearby countries, although Greek communities play a distinctive role in western Europe, North and South America, and Australia.

Demography. The Greek population has never displayed high rates of growth, although—despite grievous losses in a succession of wars and constant emigration as a result of poor economic conditions—it has usually shown a regular increase since the first census, in 1828. Most of its growth in the years since it gained its independence from the Turks resulted from two factors—annexations of surrounding areas (the Ionian Islands; Thessaly and Árta; Epirus, Macedonia, and Crete; Thrace; and the Dodecanese) and the influx of more than 1,000,000 Greek refugees from Asia Minor in the 1920s. Emigration has continued to be a limiting factor: the years 1911–15 were an active period, and emigration became particularly heavy after World War II. The most common destinations of the emigrants have been the United States, Canada, Australia, and, more recently, West Germany and Belgium.

Vital statistics are comparable to those of the developed countries of Europe. Higher birth rates after World War II have produced a youthful population. Along with population growth, urbanization has been a most important factor, especially since World War II. Most Greeks live on the main peninsula, with nearly a third living in the Greater Athens area. The rural component has shrunk to about one-third of the total. All these factors have had important social and political implications.

THE ECONOMY

Although still characterized as a developing country and handicapped by few natural resources, a low level of industrial development compared to the rest of western Europe, and a prevalence of bureaucracy and political favouritism,

Greece has achieved rapid economic growth since World War II. Although Greece is considered a Balkan country, its principal economic links are with the United States; the European Economic Community (EEC), of which it became a full member in January 1981; and the Organisation for Economic Co-operation and Development (OECD), of which it is a full member.

The Greek economy has traditionally been based predominantly on private enterprise, with the state limiting its intervention to tax incentives, cheap financing, and guarantees to overseas suppliers. Although the private sector in Greece remains considerable, the public sector has grown rapidly as a result of ever-increasing demand for government services. By the 1980s, the government had partial or total control over more than two-thirds of the economy—energy, shipbuilding, communications, transportation, insurance, and banking and, through the banks, most of industry. Shipping, however, has remained exempt.

Economic policy

During the late 1950s and early '60s, the government initiated programs to reclaim land, provide credits for agriculture, develop new farms, and improve the quality of agricultural products. The rapid rise in the production of tobacco, wheat, and cotton established the foundation for expansion in the textile and clothing industries. Other government priorities included the development of natural resources and the creation of basic industries. Several major industrial projects initiated in the 1960s included oil refineries, a modern shipyard, and an aluminum plant.

Trends in employment since the 1970s have included a substantial decline in the number of those working in agriculture and an increase in the proportion of the work force employed in industry and services. Unemployment remains relatively low. During the 1960s many Greeks had found work in the more developed countries of western Europe, particularly in West Germany. As the availability of overseas jobs diminished in the 1970s, however, many workers returned to Greece.

Resources. Greece is not rich in mineral resources. Bauxite constitutes nearly all of the country's mineral and quarry deposits, which are estimated to be more than 500,000,000 tons. There are also important reserves of lignite and chromiferous iron ore.

About four-fifths of the total mine output is accounted for by lignite production, and most of the lignite mined is used by the power stations that provide more than half of the country's electricity. The remainder is important in the manufacture of chemical fertilizers and domestic fuels. Greece is a major European producer of bauxite, much of which is shipped to the Soviet Union and Romania, although a substantial and increasingly larger portion of the ore is used in the local production of alumina and aluminum metal.

Other exploited minerals include ceramic clays, gypsum, asbestos, marble, magnesite, lead, zinc, and nickel. There are ordinary iron ores, ores with nickel content, and uranium deposits. Exploitation of the Primos oil and natural gas deposits in the northern Aegean Sea began in 1981, and a natural gas field was discovered that year off the Peloponnese.

Energy. Lignite and waterpower are Greece's most important energy resources. The Public Power Corporation was established in 1950, with the aid of the United States, to construct and operate electricity-generating plants. By 1955, four major power plants had been erected, and new power transmission and distribution lines enabled remote villages to receive electricity. The output of electricity in Greece was almost doubled between 1961 and 1971 by the building of thermal power stations, such as the one at Megalópolis in the Peloponnese, that use lignite of remarkably low calorific value and high moisture content. Several units of the Kremastá hydroelectric station began operation in 1965, but the dry climate prevents a major expansion of hydroelectric power.

Agriculture, forestry, and fishing. About 30 percent of the total area of Greece is either suitable for cultivation or actually exploited as arable land. Rough pastureland and forest compose another 60 percent of the land. Agriculture is important to the Greek economy despite the natural poverty of the rocky soil, scarce rainfall, an inheritance system that has led to excessive fragmentation of agricultural holdings (about four-fifths of which are less than 12 acres [4.8 hectares] in size), the use of inefficient farming methods, and the drift of male farm workers away from the land.

Principal crops are wheat, corn (maize), barley, oats, rice, cotton, tobacco, and potatoes, raised mostly in northern Greece, and melons, tomatoes, grapes, olives, and citrus fruits, grown in central and southern Greece and on the islands. There has been an increasing diversification of fruit crops for export, and attempts have been made to improve dairy and meat production. Sheep and goats are the main livestock, and Thessaly is the leading cattle-raising area.

Productivity has increased, although hampered by soil erosion, lack of fertilizers, and insufficient capital investment. The government has subsidized agriculture, extended liberal farm credits, and exempted agricultural income from most taxes in an effort to further expand production. Measures also have been taken to halt the trend toward ever-smaller farm holdings and to provide heavy farm equipment for hire at low prices.

Membership in the EEC has brought other changes to Greece's agricultural sector. Greece is the EEC's only producer of cotton and grows much of its fruits and vegetables and about half of its tobacco. Although subsidies from the EEC have been helpful, they were originally designed mostly for northern European farms, which are generally larger and run more like businesses than are Greek farms. Small landholdings, resistance to the formation of agricultural cooperatives, the decline in the farm population, and inefficient marketing techniques continue to plague Greek agriculture.

Forests, covering about one-fifth of the total land area, mostly belong to the government. Progress has been made in restoring the forests that were severely depleted during World War II. Forestry is not economically important—resin and turpentine are the principal products—and many forestry products are imported.

The fishing industry

The long coastline and numerous islands of Greece help nourish the fishing industry, which has grown in response to modernization of the fishing fleet and extension of refrigeration and processing facilities. On the other hand, Greece is a net importer of fish products, and the once-important sponge-diving industry, whose fleet traditionally was based in the Dodecanese, has declined in the face of competition from synthetic products.

Industry. The Greek industrial sector is characterized by a handful of large-scale companies, which either belong to the government or remain in the hands of their founding families, and thousands of smaller firms. Industry is concentrated in Athens, despite the problems of high rents, smog, and traffic congestion, and in Thessaloníki, although the government has taken steps to steer industry to other parts of the country.

Prior to the 1960s there was almost no serious industrial investment in Greece. Most of the multitude of small-scale enterprises were family-owned businesses or workshops of individual craftsmen that produced textiles, beverages, processed foods, and tobacco or were engaged in ship repair and related activities. Some expatriate shipowners began to invest their profits in Greece in the 1960s, and their investment, together with other foreign capital and legislation that provided a favourable investment climate, resulted in considerable industrial development. New projects included the building of the country's largest shipyard at Skaramanga near Athens, the aluminum plant at Dhírfis, and the huge Esso-Pappas complex (sold to the Greek government in 1981) at Dhiavatá, near Thessaloníki, consisting of an oil refinery, an ammonia plant, a steel mill, and related facilities for the production of plastics and fertilizers. The textile and cement industries, two of the oldest industries in Greece, were expanded during this period.

Although manufacturing has grown rapidly and ranks ahead of agriculture in earned income, most Greek industrial establishments employ fewer than 10 workers, and most of the raw materials, machinery and parts, and fuel they use must be imported. Some of the larger companies

(notably those that produce cement, basic metals, textiles, clothing, and plastics) have been able to avail themselves of the enlarged market provided by the EEC. These industries, however, are mostly high-cost and produce goods with little domestic value. Expectations among the Greek people for a higher standard of consumer goods than Greek industry can provide has resulted in the importation of such goods, contributing to an unfavourable trade balance.

Construction is an important industrial sector in Greece. During the late 1960s and early '70s the Greek government lifted or relaxed restrictions on lending for the construction of tourist facilities and dwellings. Tourism has continued to grow, although much of the construction of facilities has shifted from Athens to the outer islands. Dwellings account for more than two-thirds of all building construction. The Greek predilection for ownership of real property, the tradition of including a house in a girl's dowry, the use of funds sent back by Greek workers abroad to build retirement homes, and migration to urban areas have contributed to the steady pace of construction.

Greek shipping

Greek owners occupy an important place in the world of international shipping. About three-fourths of the Greek merchant fleet flies the Greek flag, and in total number of vessels, as well as in gross tonnage, the fleet is one of the largest in the world. The Greek fleet, composed of ships relatively older than those of other shipping nations, specializes in "tramping," or going anywhere on a single trip, rather than in travelling regular routes. Through a domestic tonnage tax and especially because of the money sent home by Greek sailors, Greek shipping makes a sizable contribution to the country's balance of payments, and Greek shipowners have provided much funding for the country's capital investment. Shipping provides about one-third of the national income derived from services.

Ancient ruins, beautiful beaches and islands, and abundant sunshine have made Greece a favourite destination for tourists, whose numbers annually equal about two-thirds of the population. This massive influx spawned a tourist industry that has had a very favourable effect on the country's economy, especially on its balance of payments.

Finance. The central bank is the government-controlled Bank of Greece, founded in 1928. It is the sole bank of issue and supervises all of Greece's banking operations, implementing the monetary policy decisions that are made by the government's Currency Committee. Consisting of five Cabinet ministers, this body formulates all monetary, credit, and foreign-exchange policy. As a result, monetary policy in Greece, including lending rates, is set by the government rather than by market conditions or the availability of capital. The capital market is still relatively underdeveloped. There is one stock exchange, located in Athens, in which there is active trading.

Trade. Exports have increased at a somewhat faster rate than have imports; because imports have been much greater in value, however, the trade gap has grown steadily. Crude petroleum alone accounts for one-tenth of the total value of imports. Other major imports include meat, iron and steel, passenger automobiles, milk and cream, and medicines and pharmaceuticals. An encouraging feature of trade has been the growth in exports of industrial and handicraft products, which together account for more than half of the country's total exports. Cement, textiles, and clothing remain Greece's most internationally competitive industrial products. Food—particularly dried fruit and olive oil—beverages, tobacco, and other agricultural products provide more than one-fourth of total exports. Petroleum products, minerals and ores (including bauxite), and aluminum are also important exports.

Trade with EEC members, and with West Germany and Italy in particular, is high. The negative balance of payments is alleviated by a steady surplus of invisible services (notably shipping), tourism, and funds sent back to Greece by sailors and workers who have emigrated to other countries. The International Fair of Thessaloníki is the country's principal international trade fair. (Ed.)

Transportation. The Greek landscape and seascape have greatly affected the development of the country's transportation patterns and perhaps help to account for the preeminence of Greek names in the world of international shipping. The needs of tourism, military and political considerations, and the general economy have encouraged the development of a modern national transport system.

Sea transport. As noted above, Greek and Greek-owned shipping fleets are important in national and international trade. Partly as a result, port development has been strongly emphasized. Piraeus (Piraíevs), the port of Athens, is the major centre, followed by Pátrai (Patras), Préveza, Iráklion, Kaválla, and Vólos, among others. There is a developed steamer service to the various islands, and car ferries ply many straits and inlets. Cruise ships and private vessels add to the demand for service facilities at ports.

Greek ports

Road transport. Much of the road network is unpaved, but the Athens–Pátrai, Athens–Lamía, and Athens–Thessaloníki highways are modern roads. An extensive network of rural bus routes is found on the mainland and Crete.

Railway transport. Extensive modernization has been effected in the Greek railway system, the aims being to improve the existing tracks, to standardize differing metric gauges, to forge links with western Europe, and to coordinate development with that of the roads. The Greek topography and a late start (railways date only from the 1880s, and Greece was one of the last European countries to develop them) have made this modernization costly and difficult. There are no railways on Crete.

Air transport. Air transport is operated by the government-owned Olympic Airways, which took over the financially troubled Greek National Airlines (founded 1951) in 1957. Increasing tourist traffic has resulted in a major expansion of facilities. Athens has a modern air terminal, regional facilities have been improved, and there are airstrips on Crete and some of the other islands.

(E.I.U./Ed.)

GOVERNMENT AND SOCIAL CONDITIONS

Government. Much of Greek political history in the 20th century has been characterized by civil wars, military coups, national emergencies, and other political disturbances. The constitution, introduced in 1975 after a period of military rule (1967–74), established Greece as a parliamentary republic and incorporated a number of measures designed to ensure stability in government. Among these measures was the division of the functions of the executive between a president, as head of state and supreme commander of the armed forces, and a prime minister, as head of government. In 1986, amendments to the constitution stripped many of the executive powers from the president, such as the right to dismiss the prime minister, and transferred them to Parliament.

Presidential powers, therefore, became largely ceremonial, though the president is still the supreme commander of the armed forces. The president is elected by Parliament for a five-year term and is eligible for a second term. Such indirect election ensures that no president, by claiming a direct mandate from the people, can challenge the political direction taken by the majority in Parliament. The president can dissolve Parliament only if two governments have, in rapid succession, resigned or failed to receive a vote of confidence. The 1986 amendments retained the president's power to return legislation to Parliament for reconsideration, but Parliament needs only an absolute majority to pass such returned legislation.

The prime minister, the most powerful political official in Greece, is appointed by the president. The prime minister is the leader of the party with an absolute majority in Parliament or, if no such party exists, the leader of the party with the largest number of seats who is capable of forming a government. The prime minister has the power to dissolve Parliament.

The Vouli

Parliament (Vouli) is a unicameral legislative body composed of 300 deputies chosen by direct, universal, and secret ballot for four-year terms. Parliament elects its own speaker, or president, must meet once a year, has the power to impeach the national president and other members of the government and to declare a state of emergency, and may revise the constitution.

Electoral laws are not specified in the constitution, and Greek electoral procedures have therefore varied, each government trying to modify the appropriate laws to its

advantage. Voting is mandatory for all Greeks aged 18 years or older.

Greek politics is distinguished by a highly personal approach, not only in the influence of national leaders but also in the traditional patronage system grounded in exchanges of favours. This system has resulted in a largely inefficient government bureaucracy and in political parties traditionally organized around persons rather than policies.

There were four major political parties in the late 1980s. The Democratic Centre Union (originally called the Centre Union–New Forces Party) was formed through a merger of the Centre Union and the Movement of New Political Forces. The Centre Union represented the more conservative wing of the Greek liberal movement, the Centre Union Party founded in 1961 by Georgios Papandreou. The Movement of New Political Forces was founded to campaign for policies of democratic socialism.

The New Democracy Party was founded by Konstantinos Karamanlis. It pledged to work democracy in Greece through political, economic, and social reform.

The Pan-Hellenic Socialist Movement (Pasok), incorporating two resistance organizations, was founded by Andreas Papandreou, son of Georgios Papandreou. Its platform called for a nonaligned foreign policy and nationalization of private enterprises. Pasok is the first Greek political party with mass appeal and an organization based on mass membership.

The United Left Alliance was formed as a union of the United Democratic Left (Eneia Dimokratika Aristeras [EDA]; founded 1951) and the two factions of the Communist Party of Greece (Kommunistikon Komma Ellados [KKE]). The Communist factions maintain separate organizations as the KKE-Exterior, which supports the policies of the Soviet Union, and the KKE-Interior, which does not acknowledge the preeminence of the Soviet Communist Party.

Although democracy was restored in 1974 and officers returned to their military duties, Greek politicians continue to take into account the possible responses of the military forces to major governmental decisions. The Greek armed forces consist of an army, navy, and air force recruited on the basis of compulsory military service for all male citizens aged 21 for a period of two years. The National Guard is recruited from reservists for local security duties in frontier districts.

The city police and the gendarmerie, the latter a paramilitary force responsible for the whole country outside Athens and several other cities, were reorganized in 1984 into a single unit, the Greek police. It is administered by the minister of the interior.

The judiciary

The three most important court systems in Greece are the Supreme Court, State Council, and Comptroller's Council. The Supreme Court consists of both civil and criminal sections.

Below the Supreme Court, courts of appeal have jurisdiction in cases of criminal and civil law of second degree. Below these are the courts of first instance, which function also as criminal courts. Criminal and juvenile courts sit in towns with courts of first instance. Some towns also have tax courts. There are courts of justices of the peace and magistrates' (police) courts throughout Greece.

Procedure in the courts follows the French model; *i.e.*, the investigating magistrate examines the evidence and interrogates witnesses and, if he decides that there is a prima facie case, refers it to the public prosecutor, who decides whether or not a charge will be brought. Judges of the higher courts are appointed for life and others, appointed for fixed terms, may be removed only if convicted of a criminal offense.

Greece is divided administratively into 13 *dhiamerísmata* (regions). The country is further subdivided into more than 50 *nomoí* (departments; singular *nomós*), each administered by a *nomarch* who is nominally appointed by the minister of the interior. Special arrangements are provided for the Greater Athens area, and the peninsula of Mt. Athos in Macedonia is a self-governing monastic community with a civil governor, appointed by the government, who is responsible for public order outside the monasteries.

Municipal government, with elected mayors and urban and rural councils, was established in Greece in the mid-19th century. Local authorities may levy certain taxation, but, in general, provincial services are supported by the national government.

Education. Education is highly regarded in Greece, partly because of the pride the people have in the Greek classical heritage and also because of the traditional value of the university degree as a means of achieving upward social mobility. The centralized, state-run school system has been extremely successful in reducing illiteracy, despite a shortage of teachers and schools. The Ministry of Education and Religion administers the system, and funding is provided by the central and local governments. Private schools, particularly numerous at the secondary level, are regulated by the state.

Compulsory schooling was first introduced in 1929. Educational reforms have increased the number of years of compulsory education from six to nine, required the use of demotic Greek (the vernacular language) rather than Katharevusa (the literary language) in the schools, and provided free education at all levels.

Higher education

The curriculum of higher education has been concentrated in law, philosophy, and the humanities, with the result that many Greek students seeking degrees in scientific and technical fields must attend foreign universities. The government also has limited the number of openings at Greek universities in anticipation of the number of job openings for graduates. For higher studies there are universities at Athens (National Capodistrian University, founded in 1837, and National Technical University, founded in 1836) and Thessaloníki (Aristotelian University, founded in 1925), as well as the University of Ioánnina (1964), the University of Pátrai (1964), the University of Thrace (1973), and the University of Crete (1973). Athens and Thessaloníki also have a number of higher schools with university status, each specializing in a separate field.

Private universities are banned by the constitution. The University Council advises the government on planning for higher education, and the Academy of Letters and Sciences sets and implements university standards.

Health and welfare. After World War II the government took the lead in combatting disease by establishing modern services and facilities. Malaria, pulmonary tuberculosis, and dysentery, once major scourges, have been controlled. Standards of hygiene and sanitation have been greatly improved, the nation's water supply is generally safe, and the incidence of typhoid, formerly of epidemic proportions, has been dramatically reduced.

The Ministry of Health and Welfare is responsible for the maintenance of hygiene, malaria control, and the establishment and financing of health and medical care centres and pharmacies, as well as for providing drugs and maternal, child, and psychiatric care. Health care is provided free to rural inhabitants if they are unable to pay. The large hospitals are concentrated in Athens, Thessaloníki, and Pátrai. Most of Greece's complex social insurance schemes are controlled and supervised by the Ministry of Social Security.

Drastic changes were made to Greek family law in the early 1980s. Civil marriages were permitted for the first time, dowries as a legal requirement for marriage were abolished, equality was guaranteed between marriage partners, and the naming and registration of children with the civil authorities, rather than with the church, was allowed. In addition, divorce was made easier. (Ed.)

CULTURAL LIFE

The physical remains of the culture of ancient Greece, whether preserved in situ, in the fine network of museums, or in the museums of other countries, are constant reminders of the country's classical heritage. It continues as an important element in the culture of contemporary Greece and plays a major role, through its attraction to tourists, in the economy. In addition, the deep religious traditions of the country—which found rich expression in the medieval icons and in the mosaics and frescoes that made the 14th century one of the remarkable eras of Byzantine art—continue to provide a fertile cultural

source, generating a great variety of contemporary folk art and religious festivals. These are at their most vigorous in the rural areas generally and in the more remote regions in particular. Also remarkable and influential was the century from 1560 to 1660, when the arts flowered on Crete. Literature of the period included *Erotókritos,* a romantic epic that has become a national poem for some Greeks. The most famous Cretan artist of the period was Doménikos Theotokópoulos, who gained universal fame as El Greco.

Easter is the major event in the Orthodox calendar, and the sombre processions of Good Friday are followed by the festivities—including roasting of lambs and dancing in traditional costume—celebrating Easter Sunday. Other religious ceremonies occur throughout the year, beginning with the New Year blessing of the sea at Piraeus and elsewhere. The summer months are renowned for international music and drama festivals, particularly at Athens and Palaiá Epídhavros (Epidaurus). A revival of Byzantine iconography may be seen in the work of such artists as Fotos Contoglou and Stathis Trahanatzis as well as in the restoration of medieval frescoes and mosaics. The traditional Karaghiozis puppet theatre has been preserved by such masters as Panayioti Mikhopoulos.

Modern Greek writers

Modern Greek poets are considered by many to be among the best of the 20th century. Poets of international renown include Constantine Cavafy (who spent most of his life in Egypt), George Seferis (awarded the Nobel Prize for Literature for 1963), Angelos Sikelianos, Odysseus Elytis (awarded the Nobel Prize for Literature for 1979), and Yannis Ritsos. Another remarkable writer was the Cretan Nikos Kazantzakis, who experimented with several forms of literature. Contemporary Greek literature is fortunate in its English translators, especially the Americans Kimon Friar, Edmund Keeley, Themis Vasils, and Theodora Vasils.

Greek musicians with international followings include Manos Hadjidakis and Mikis Theodorakis. After the fall of the military regime in 1974 there was a revival in Athens of an irreverent political theatre. The film industry was also revived.

The Greek character is reflected in the opening verses of the "Hymn to Liberty" written by the 19th-century poet Dhionísios Solomós. These lines, set to the stirring music of N. Manzaros as the national anthem of Greece, read (in a prose translation by Theodora Vasils):

> I know you by the sword's dread cutting edge, I know you by the look that with vigour measures the earth. Out of the sacred bones of the Hellenes you issue valiant as before, hail o hail, Liberty! You dwelt therein, sorrowful, withdrawn, waiting for a mouth to tell you, "Come again." That day was long in coming, and everything was silent, cowering under the terror and the crush of slavery.

The other great "national" song of the Greeks is the fervent Resurrection hymn attributed to an 8th-century monk, St. John of Damascus,

> Christ has risen from the dead, by death trampling upon death, and has upon those in the tombs life bestowed.

The music of these two hymns, a millennium apart in composition, is both joyful and haunting. The listener is reminded of Greek nostalgia, of the deepest fears and hopes of the Greeks and indeed of all mankind.

For statistical data on the land and people of Greece, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (G.An.)

History

The peoples of what is now Greece created the first enduring civilization in Europe, one that is the wellspring of Western culture. The Minoan civilization—the first great civilization in Europe—flourished on Crete during the Bronze Age and was later absorbed into the Mycenaean culture of the mainland. Migrations and invasions of the mainland, over thousands of years, by peoples from Asia Minor and southern Russia created a population that finally merged into one people. Conscious of their unity in language, religion, and culture, they called themselves Hellenes and their homeland Hellas. They created art, architecture, literature, philosophy, and scientific thought that have never been surpassed, but they did not develop political unity. Barriers of geography, interest, and natural temperament divided the Hellenes uncompromisingly.

EARLY CULTURES

The recorded history of Greece begins about 800 BC, but significant archaeological remains exist from the New Stone Age, before 6000 BC. Using stone tools, the earliest known inhabitants cultivated the vine and olive, wheat and barley, peas and beans, and kept flocks. They made fine pottery and did not fortify their small settlements. They worshipped the fat Mediterranean fertility goddess and may have lived without fear of invasion for more than 3,000 years. This was, according to Greek legend, the golden age.

The Bronze Age. The centuries from the beginning of the Bronze Age in mainland Greece, about 3000–2200 BC, to the end of the Mycenaean period, about 1100 BC, are called the Helladic period. During that time, the culture changed from that of a Middle Eastern people living a simple pastoral and agricultural life to one of Indo-European-dominated principalities ruled by warlike chieftains engaged in perpetual conflicts with each other. The princes profited from maritime trade throughout the islands and along the coast of the eastern Mediterranean. Their power was based on the tenure of land, which they acquired for their inferiors to work. They spent their lives in combat and the pursuit of everlasting fame.

Such changes in Greece came slowly. An invasion about 2600 BC from Asia Minor brought people into the land who knew metalworking, improved agricultural practices, and sailing. About 600 years later Indo-European tribes, part of a westward migration, poured into the peninsula from the north and overturned the existing society with terrible destruction. They spoke an Indo-European language and were of a different racial stock; and their beehive-shaped royal tombs were distinctive. In time, they acquired the skills of the previous inhabitants but lived in fortified complexes, each built around a great hall surrounded by columns. Ramps, staircases, and a series of vestibules led from the city gate directly to the hall (the megaron), where the sacred fire burned. This plan eventually became that of the Greek temple.

Mycenaean period

When the mainland rulers encountered the Minoan civilization of Crete about 1600 BC, the Mycenaean culture (named for the mainland city of Mycenae) resulted from the mixture. The Minoan civilization on Crete supplanted the Stone Age trading culture of the Cyclades Islands. With its overseas trade, palaces, urban life, and artistic achievements (including writing), the Minoan is considered the first true civilization in Europe. The Minoan monarchy eventually became subject to Mycenaean military power; the city of Mycenae, in turn, so much valued the Minoan artistic genius that it imported Cretan artists. Thus, Greek sculpture began on Crete, where Minoan wall paintings still arouse admiration.

Mycenae dominated the other cities of the period—Pylos, Tiryns, Thebes, Orchomenus, and Athens—its wealth confirmed by shaft graves of about 1600 BC. According to Greek legend, Agamemnon, the king of Mycenae, directed an all-Greek expedition against the rich city of Illium (Troy), near the Dardanelles. The 10-year war (*c.* 1200 BC) ended in Illium's downfall but devastated the Mycenaeans as well.

The Dorian invasions. During the 12th century BC, a catastrophe attributed to the Dorian invasion reduced the Mycenaean palaces to ruins and overwhelmed the people who did not escape to Athens, the islands, or Asia Minor. The Dorians had a military society, used iron weapons, and worshipped male, more often than female, gods. They took over the old shrines and rededicated them but did not rebuild the megarons and cities. Their way of life was so rude that the next 300 years are sometimes called the "Dark Ages"; they are also called the Geometric period for their simply decorated pottery.

As Greece began to recover from the effects of the invasions, the Greek people developed a language and religion in common with the Dorians, and the populations began

The Olympian gods

to assimilate. Everyone acknowledged a family of gods called the Olympians, who were said to live in palaces on Mt. Olympus—gods with the same names as earlier ones but with somewhat different manifestations. Their believers worshipped with seasonal festivals, athletic contests among cities, and observances dedicated to the protective deity of each city. The best known of these religious celebrations was the Olympic Games, held every four years at Olympia in honour of Zeus and Hera. The games began in 776 BC, the first recorded date in Greek history. Thereafter, the Greeks dated events by reference to the year of the Olympiad.

THE ARCHAIC PERIOD

The terror of the Dorian invasions resulted in the formation of city-states, the little nations of the Greeks. People from scattered villages moved near Mycenaean fortified places for protection; from these beginnings, often remembered in legends of heroic founders, the city-state evolved. From about 800 BC, their design followed a similar plan. A high-walled citadel (the acropolis) contained the temples and could accommodate the population and people from nearby villages when the city was under siege. Below were the market area (the agora) and the residential quarters. City-states had independent governments, definite boundaries, and diplomatic relations. Greek political history is largely an account of five major city-states—Athens, Sparta, Thebes, Corinth, and Argos.

The Greek colonies

The Archaic period proper began about the middle of the 8th century BC and extended to the beginning of Classical times, about 500 BC. Spurred by population growth on the Greek mainland, colonists established new cities from Asia Minor and the Black Sea to France, Spain, and North Africa in the great colonizing movement. Refugees from Athens established the first colonies in Asia Minor, aided by Lydia, a rich inland state there. The Ionian cities throve from the Black Sea trade. All inhabitants of new cities in Asia and around the Mediterranean considered themselves Greeks among barbarians; they maintained ties with the mother cities that sent them and successfully competed for markets with the Phoenicians. (Some of these cities—Byzantium [Istanbul], for example—still exist.) In the late 7th century BC, commerce was revolutionized by the use of coinage, which the Ionian Greeks learned from the Lydians.

What is considered Greek literature and philosophy first appeared in the Ionian cities. The Greek alphabet was created there, and Homer composed the first poetry in Greek—the *Iliad* and the *Odyssey*—based on epics about the Trojan War and the wanderings of Odysseus. The Greeks considered Sappho of Lesbos the best of the lyric poets, and Thales of Miletus emerged as the first Greek philosopher. On the Greek mainland, in Boeotia, Hesiod wrote during this period.

Influenced by Asia Minor and Egypt, Greek art took new forms. Vases decorated with figures and scenes from mythology replaced those with geometric decoration. Athens, with its superior potters and painters, captured the market and created the famous Attic style. Sculptors began to carve in stone and marble, and leading families placed statues of gods and young aristocrats in temple areas. The statues were rigidly posed but graceful and often wore the mysterious "Archaic smile." Temples, decorated with carved reliefs, took their final architectural form. Thus, the foundations of Classical Greece were laid.

THE CLASSICAL PERIOD

The 5th century BC, the beginning of the Classical period, was both desperate and glorious for the mainland Greeks. The Persians invaded their land twice, at devastating cost. Yet it was also a century of victory and the zenith of Athenian culture.

Greco-Persian wars

The Greco-Persian wars began in 499 BC when Ionian Greek cities revolted against rule by the Persian Empire, and Athens aided them. The Persians crushed the rebellion and Darius I the Great resolved to punish Athens. In 490 BC he launched an invasion force, which the greatly outnumbered Athenian army, with its Plataean allies, drove back to the sea at Marathon. Ten years later, Xerxes of Persia led an enormous expedition of revenge. The Persians, after killing all the defenders of the pass at Thermopylae, destroyed Athens and burned the Acropolis. However, Themistocles, commander of the Athenian fleet, led the Greek ships that destroyed the Persian navy in the Bay of Salamis. Without naval support, the Persian army was soundly defeated at Plataea in 479 BC by a confederation of Greek cities. The Greeks' successful defense of their independence against the much more powerful Persian Empire was due to the courage of despair and their love of liberty. The victory gave Athenians the self-confidence to rebuild their city on a grander scale and to establish an empire of their own.

By this time Athens and Sparta, the two leading city-states, were contrasts in government and principles. Sparta, by tradition under its lawgiver Lycurgus, had instituted an authoritarian military regime and was ruled by two kings, their council, and a few thousand citizen-soldiers. Spartans emphasized hard living and rigid adherence to tradition. Men were trained to win wars and women to rear brave sons. Sparta avoided ventures far from home, fearing the revolt of neighbouring cities or of the helots (slaves).

Solon, the lawgiver of Athens, provided his city with a constitution that allowed a broader base of representation in government. Solon's plan, through time and despite interruptions, led to a government that, in the late 5th century BC, made its laws in an assembly of all citizens, rich and poor. All votes were equal; the quorum on some questions was more than 6,000. The system would have been too cumbersome if the city had not had inspired leaders who could win support and to whom great powers would be delegated by the assembly. Athens prospered under democratic government, especially under the leadership of Pericles from 460 to 429 BC.

The Delian League

As Greece, under Athenian leadership, freed the Aegean Islands and Ionia from Persian rule, Athens insisted on a confederacy for protection against Persia. This confederacy, with the levies from each city that composed it, was based on Delos. The transfer of that treasury to Athens in 454 BC clearly indicates that the Delian League had become an Athenian empire. Piraeus had become the leading port in the Mediterranean, and Athens, wealthy from tribute and commerce, had the resources to fulfill the ambition of Pericles to make the city "the school of Greece." Under Pericles the drama, as well as public works, flourished.

Sparta, upon seeing the imperious role Athens had assumed in Greece, grew increasingly fearful. Other important cities, such as Corinth and Thebes, joined Sparta in opposition that led to the Great Peloponnesian War (431–404 BC). Thucydides described the war in his masterly history, perceiving from the outset that this conflict would be forever important as one in which fundamental issues were at stake. The war ended in defeat for Athens. The irreplaceable Pericles had died of the plague that devastated Athens in the first year of the war; and an ambitious assault on Syracuse, a Sicilian city allied with Sparta, failed with the loss of almost the entire expedition. Defeats on sea and land finished Athenian resistance.

Greek cities that had thrown off Athenian rule soon discovered that Spartan domination was even more harsh. A long-feared revolt of the helots, however, markedly reduced Spartan power abroad. Thebes then held a brief supremacy; afterward, Athens began to recover its power.

Athens was preeminent in Greek intellectual life, even in hard years. Socrates, the great Athenian philosopher, questioned the assumptions on which the city operated and opened new fields of thought. He was executed in 399 BC for "impiety" and other infractions of Athenian law. His disciple, Plato, immortalized Socrates in his dialogues. Xenophon, another disciple, also wrote about Socrates but is remembered today more for his histories than for philosophy. Aristotle, having learned much from Plato, dedicated his phenomenal capacities to investigations and considerations of natural history, poetry, ethics, and government.

Philip II of Macedon

In 359 BC the undeveloped land of Macedonia, on the northern frontier of Greece, acquired a new king in Philip II. He was a military genius and a skillful diplomat who as

a youth had spent three years as a hostage in Thebes. He united the warring clans of his own country and reorganized its army, using Theban models. By manipulation of the divided cities of Greece and by conquest, he expanded Macedonian power. The Athenian orator Demosthenes persuaded the factions of his city to resist Philip, but Athens and Thebes lost the decisive Battle of Chaeronea in 338 BC. The victorious Philip united all the Greek cities but Sparta and, by appointing himself captain-general of the confederation, he made Macedonia the true ruler of Greece.

Philip was assassinated in 336 BC in the midst of preparations to invade Persia. He was succeeded by his son Alexander, who was then 20 years of age. Alexander ruthlessly put down the tribes and cities that considered the time ripe for rebellion. In 334 BC he crossed into Asia, successfully challenging Darius III. Alexander and his general staff, known as the Companions, led the Greek armies across Asia to the Indus River Valley in India. He conceived a plan for an empire that would unite Greeks and Persians into one people and one civilization, but he died of fever at Babylon in 323 BC. (For a more detailed treatment of early Greek history, see GRECO-ROMAN CIVILIZATION, CLASSICAL.) (S.P.A.)

MACEDONIAN PERIOD AFTER ALEXANDER

Macedonia and the Greek city-states, 323–224 BC. The conquests of Alexander III the Great shifted the centres of political power from Greece to the Middle East and Egypt. The Diadochi (Successors) established dynasties in cities such as Pergamum, Antioch, Seleucia on the Tigris, and Alexandria in Egypt and encouraged commerce, art, and learning. In Alexander's former empire, Greek traditions and language were the possessions of the foreign ruling class, and literature and science became restricted to a small circle around the court. Royal patrons established large libraries in Pergamum and Alexandria, the latter coming to hold more than 500,000 scrolls. The Ptolemies of Egypt founded the Museum in connection with the library to make it possible for scholars to live and work in Alexandria. For more than 500 years the Museum was a centre of study for scholars, including the mathematicians Euclid, Apollonius of Perga, and Archimedes, the astronomers Eratosthenes, Aristarchus of Samos, and Ptolemy, and the physician and anatomist Galen.

Library at Alexandria

Hellenistic literature began with New Comedy, based on middle-class family life. Its chief playwright was Menander, whose influence was extensive in his time, though very little of his work has been recovered. The best poetry, composed in Alexandria, was intellectual, written for a sophisticated urban audience. The chief poets were Theocritus, Callimachus, and Apollonius of Rhodes.

Sculpture was the popular art. It has been said that at times the statues in a Hellenistic city outnumbered its inhabitants. Poses and expressions showed more obvious emotions, as artists worked to convey their personal conceptions. Fine portrait busts and masterpieces such as the Venus de Milo and the Victory of Samothrace are Hellenistic creations. Also characteristic of the Hellenistic spirit were giant sculptures such as the 100-foot-tall bronze figure of the Colossos at the harbour of Rhodes and the great altar of Zeus at Pergamum.

Despite these developments in other places, Athens remained supreme in Classical art and philosophy. Philosophic schools, among them the Stoics and Epicureans, emphasized individual character and actions, while religion became less civic and more concerned with personal salvation. The older mystery cults in Greece grew, and foreign deities—Isis and Serapis, for example—gained worshippers. As political fortune seemed to turn against their cities, the European Greeks began losing faith in the Olympian gods upon whom they had relied as guardians of their communities.

The wealthy class in most Greek cities favoured good relations with Macedonia. The democratic leaders were afraid of a coup d'état by the wealthy and resented the restrictions on foreign policy that had been imposed by the League of Corinth, an alliance of all Greek cities but Sparta that had been organized in 337 BC by Philip II of Macedon. As long as Alexander had lived and conquered, they complied reluctantly with his wishes. But the news in 323 BC of his death and of the quarrels among his generals gave the democratic leaders some hope of breaking the alliance with Macedonia. Large forces of Greek mercenaries were under arms: 8,000 men, commanded by an Athenian, Leosthenes, were at Taenarum in Laconia, and 23,000 men were marching homeward from the east. The Athenian people decided to go to war. They engaged Leosthenes and his mercenaries and formed an alliance with the Aetolian League, but most of the Greek cities—especially those that had reason to fear Athenian ambitions—opposed this policy or at least refused to cooperate. The allies began the Greek War, as they called it (it is also called the Lamian War), with an army of about 30,000 men. They seized Thermopylae in October 323 BC and blockaded a Macedonian army under Antipaterin Lamía throughout the winter. Meanwhile, an Athenian fleet of 170 ships sailed to the Hellespont but gained no allies. In 322 three Macedonian armies were operating in Thessaly, and two Macedonian fleets held the northern Aegean. Superior numbers enabled Antipater, regent of Macedonia, to defeat and split the Greek allies. Athens surrendered unconditionally in September 322. A year later the Aetolian League accepted moderate terms from Antipater, who was anxious to settle affairs and to intervene in Asia.

Antipater's victory over Greek allies

The lessons of the Greek War were clear. Athens was unable to unite the Greek cities because of their mistrust, and the forces of Athens and Aetolia alone were far inferior to the massive armies of the Macedonians. Antipater abandoned the policy of cooperation. Treating Athens as a traitor, he exacted an indemnity for the war, placed a garrison in Munychia to control the Piraeus port and Athens, and disfranchised the poorer class, so as to leave only 9,000 citizens as the electorate for an oligarchic government. The assembly sentenced to death the democratic leaders who had instigated the war. Demosthenes, the Athenian orator and statesman, committed suicide. Antipater set up pro-Macedonian oligarchies in the Peloponnesian cities and installed a Corinthian, Dinarchus (perhaps the speech writer of the same name), to act as Macedonian governor of the Peloponnese.

From 320 to 275 BC the cities of the mainland were involved in the wars of the Diadochi, who fought for power and ended by dividing the Macedonian Empire into a number of fairly stable monarchies. During these troubled years the Greek cities were important to the rival kings and generals because they were able to provide troops and taxes. The policy of cooperation was applied for the last time in 303 by the Macedonian general Demetrius I (Poliorcetes), who revived the League of Corinth and treated the Greek cities as allies. Otherwise the Macedonians kept control by setting up puppet governments and imposing garrisons. In Athens, for instance, Antipater's son Cassander in 317 placed Demetrius Phalereus at the head of an oligarchic government under the protection of a Macedonian garrison; and this situation lasted until 307. The Areopagus Council received many of its traditional powers, the rich were freed from the liturgies (public duties) imposed by the democracy, and a strong control was set on all legislation; but behind this facade Demetrius Phalereus was a tyrant who ruled with Macedonian support. The city that made the boldest bid for independence was Sparta, whose successful revolt in 280 was joined by Elis and by some cities in Arcadia and Achaea but not by rivals of Sparta in the Peloponnese. But the very success of this revolt made Sparta's allies suspicious and uneasy, and they fell away when Sparta attacked the Aetolian League and suffered heavy losses.

In 275 the Antigonid dynasty was finally established on the Macedonian throne. Between 275 and 224 the strength of Macedonia in Europe was weakened by the attacks of Pyrrhus of Epirus (died 272) and by the activities of Ptolemy II Philadelphus of Egypt and other Hellenistic monarchs. The Greek cities again obtained a greater measure of independence. The most vigorous peoples were those of the hill country who developed the federal system. The Aetolian League had been organized as a league of cantons with a democratic assembly on the model of

the Boeotian League since 367 at latest, and it had never been subjugated by Macedonia. It had been foremost in driving off the formidable Celtic invasion of 279. Now it expanded its territory by incorporating other cities, which retained their own citizenship but assumed Aetolian citizenship, and it granted an exchange of citizenships with more distant cities. The Aetolian Assembly, which met twice a year, decided on policy and elected for the year a general who presided over the assembly and commanded its armed forces; and routine administration was carried out by a council elected by the cantons in numbers proportionate to their military strength. The general and his advisers, acting as a committee, were entrusted with wide powers because the assembly met so rarely; and a man could be reelected to the generalship only after an interval of some years, lest he become too powerful and autocratic. By 224 the Aetolian League, having defeated its rival, the Boeotian League, included within its citizenship the core of central Greece from coast to coast and counted among its close allies by isopolity such states as Cephallenia and Chios. Like the democratic city-states of the previous century the Aetolian League was aggressive and warlike.

The Achaean League had been disbanded in the period of Macedonian domination. It formed again on a democratic basis *c.* 280 BC, with an assembly of citizens over 30 years of age and a representative council of deputies. From *c.* 200 BC the assembly's powers declined. It met jointly with the council at irregular intervals to decide major issues. The predominance of the council made the policy of the Achaean League less adventurous than that of the Aetolian, and the member cities had a greater degree of independence, rarely exchanging their citizenship. The number of annual generals of the league was reduced from two to one in 255. Then the general acted as president of the assembly and commander of its forces and shared executive authority with a small committee of officials; the same man could be elected general in alternate years. This feature in the league's democratic constitution enabled Aratus of Sicyon to hold office every other year and gain a strong position of leadership. Aratus had a fanatical hatred of the tyrants whom the Macedonians supported, and he drove them out of city after city until by 228 the Achaean League comprised all cities in the Peloponnese except Elis, Messenia, and Sparta.

The greatest city-states of the past led a rising against Macedonia in 267 or 266 BC. This was the Chremonidean War (so named for the Athenian leader), in which Athens and Sparta, in alliance with Ptolemy II Philadelphus of Egypt, proclaimed themselves champions of the liberty of the Greek cities. But few of the cities joined them. The Athenians fought gallantly but were starved into surrender by the superior forces of Macedonia in 262. They did not become independent again until 229; after this they preserved their freedom by a policy of neutrality.

Chremonidean War

The Spartans lost their allies and remained a second-rate power, without Macedonian intervention, until 227 BC, when they saw themselves threatened by the expanding Achaean League. Sparta indeed had been crippled since the Battle of Leuctra by the small number of Spartiates, or full citizens, who owned sufficient land to pay their contributions to the messes. In 227, however, a Spartan king, Cleomenes III, changed the situation by revolutionary methods: he abolished debts and confiscated all land, which he redistributed in lots to increase the citizen body. Having thus enlarged his army, and announcing a program of social revolution, Cleomenes attacked the Achaean League and detached so many of its members that its general, Aratus, called on Antigonus Doson, king of Macedonia, for help in 224.

Coming of the Romans. Between 224 and 205 BC the rulers of Macedonia struggled to regain a dominant position in Greece. As the price of help to Aratus, Antigonus Doson took over Corinth and founded another Greek league. This was similar to Philip II's League of Corinth and included the leagues of Thessaly, Phocis, Boeotia, Euboea, Epirus, Acarnania, and Achaea. As the elected commander of the league's forces, Antigonus cornered and defeated the Spartans at Sellasia in 222, entered Sparta, and cancelled the measures of Cleomenes. When Antigonus died in 221, Philip V succeeded to the throne of Macedonia. He helped the members of the new Greek league against the Aetolian League, Elis, and Sparta in another so-called Social War (220–217), which was indecisive. At the peace conference the Aetolian delegate spoke of "the cloud rising in the west," namely Rome. When the outbreak of the Second Punic War was followed by a series of Roman defeats, Philip allied himself with Carthage (215) and tried in vain to oust Rome from the Illyrian coast. Rome raised Philip's Greek enemies—the Aetolian League, Elis, Sparta, and Messenia—against him in 212 or 211; fighting ended in Greece in 205 without any clear gain for either side.

The period from 205 to 146 was marked by the advance of Rome. Some Greek cities called upon Rome for help, as they had called upon Persia and Macedonia, but they learned to their cost that the Romans had neither gratitude nor scruple where their desire for power was involved. The Romans' first step was to defeat Philip V at Cynoscephalae (197) and to confine him to Macedonia; they were helped in this by the Aetolian and Achaean leagues, Athens, Rhodes, and Pergamum. It was then the Romans, not the Macedonians, who proclaimed that the Greek cities were to be independent. They used Greek and Roman troops to enforce the transfer of Argos from Sparta to the Achaean League. Rome also gave the Aetolian League less territory than it claimed. The Roman army left Greek soil in 194, but the Aetolian League countered by inviting the Seleucid king Antiochus III into Greece. Antiochus came in 192, but the Romans and Philip V of Macedonia drove him out and reduced the Aetolian League to impotence. The Romans then defeated Antiochus at Magnesia, on the other side of the Aegean, in winter 190/189, deprived him of Asia Minor west of the Taurus Mountains, and announced the freedom of the Greek cities there in 188.

Advance of Rome

Meanwhile Rome's ally, the Achaean League, led by Philopoemen, was extending its territory; Sparta, which had been brought into the league in 192, tried to secede; but Philopoemen was able to enforce the annexation of the city in 188, after a number of leading citizens had been massacred. A few years of uneasy peace were marked by complaints that the Greek cities made to Rome against one another and against Macedonia. Then, in 171, the Romans ordered Perseus, the successor of Philip V of Macedonia, to disband his forces. Perseus refused and Rome attacked. The Greek cities played little part in the war. They watched the almost complete annihilation of the Macedonian army in the Battle of Pydna in 168 and the dismemberment into four republics of the kingdom that had towered over them for two centuries. The Greeks too were given a taste of Roman methods: in Epirus, where sympathy with Macedonia had been shown in two wars, the Roman army destroyed 70 towns and deported 150,000 men as slaves; although the Achaean League had declared for Rome, 1,000 leading citizens were deported to Italy; and Aetolia was weakened by Roman troops, which helped the pro-Roman party in a civil strife. Even so, the authority of Rome was flouted. Then, in 148 a Roman army annexed Macedonia as a province; and in 146 another army defeated the forces of the Achaean League—Boeotia, Euboea, Phocis, and Locris. This time the consul L. Mummius made an example of Corinth: the men were massacred, the women and children were enslaved, and the city was razed. Rome ordered the disbanding of all leagues and the replacement of democracy by oligarchy in a number of cities and announced that any breach of the Roman regulations would be punished by death.

POSTCLASSICAL HISTORY

Greece under the Roman Republic. After the collapse of the Achaean League, the Roman Senate appointed a commission to reorganize Greece as a dependency of Rome. Corinth, the chief centre of resistance, was destroyed; the national and cantonal federations were dissolved, commercial intercourse between cities was restricted, and the government was transferred from the democracies to the propertied classes, whose interests were bound to Roman supremacy. Some favoured cities, such as Athens and Sparta, retained their rights as *civitates liberae;* the

other cities continued to enjoy local self-government but probably were subjected to payment of tribute. Powers of supervision were entrusted to the governor of Macedonia. Internal disorder that remained from the previous revolutions was checked by the historian Polybius, whom the Senate deputed to mediate between the litigants.

Greece was seriously disturbed during the First Mithradatic War (88–85 BC). Many of its cities sided with Mithradates, whose success in detaching the Greeks from Rome is to be explained partly by the way in which his agents incited the imperialistic ambitions of cities such as Athens and partly by his promises of support to the democratic parties. The result of the war was disastrous to Greece. Apart from the confiscations and exactions by which the Roman general L. Cornelius Sulla punished the disloyal communities, the extensive and protracted campaigns left central Greece in a ruinous condition.

During the last decades of the Roman Republic, oppressive exactions by officials were not unknown. Still greater was the suffering produced by the rapacity of Roman traders and capitalists. Sicyon had to sell its most cherished art treasures in order to satisfy its creditors. A further hindrance to Greek prosperity was the diversion of trade that followed upon the establishment of direct communication between Italy and the Middle East. The coastal districts and islands suffered considerably from swarms of pirates who freely plundered the chief trading places and sanctuaries; *e.g.,* Delos in 69 BC. These raids came to an end with the suppression of piracy in the Mediterranean by Pompey (67 BC).

Depletion of resources under the Romans

The Greeks provided part of Pompey's excellent fleet during his conflict with Julius Caesar. In 48 BC the decisive campaign was fought on Greek soil, and the resources of the land were severely taxed by the requisitions of both armies. Though Caesar treated the country quite leniently after his victory at Pharsalus, the Greeks supported the cause of Brutus after Caesar's assassination; but they were too weak to render much service. They subsequently passed into the hands of Mark Antony, who imposed further exactions to defray the cost of his wars. The levies and requisitions that Antony made in 31 BC for his campaign against Octavian (the future Augustus) exhausted the country so much that, after the Battle of Actium, Octavian had to take prompt measures to avert a general famine. The depopulation that resulted from the civil wars was partly remedied by the settlement of colonists at Corinth and Patrae by Julius Caesar and Octavian; on the other hand, the foundation of Nicopolis Actia by Octavian merely transferred the people from the country to the city.

Roman imperial rule: Augustus to Diocletian. In his reorganization of the provinces of the Roman Empire, Augustus incorporated Thessaly with Macedonia and converted the rest of Greece into the province of Achaea under the control of a senatorial proconsul resident at Corinth. Several communities, including Athens and Sparta, retained their rights as free cities. The provincials were encouraged to send delegates to a synod at Argos to consider the general interests of the country. The Delphic Amphictyony was revived and extended so as to represent in a similar fashion northern and central Greece.

Economic conditions did not greatly improve under the empire. Although new industries sprang up to meet the needs of Roman luxury, and Greek marble, textiles, and table delicacies were in demand, the only cities that regained a flourishing trade were the partly Italian communities of Corinth and Patrae. Commerce languished in general, and the soil was mainly abandoned to pasturage. Such wealth as remained was amassed in the hands of a few great landowners and capitalists; the middle class continued to dwindle, and many persons became dependent on doles and largesses.

Seeing no future before them, the European Greeks were content to dwell in contemplation amid the glories of the past; and national pride was fostered by the undisguised respect with which leading Romans treated Hellenic culture. To perpetuate this culture, the Greeks continued to set great store by classical education, and in Athens they possessed the chief university in the Roman Empire. At the same time, the Greeks had so far lost their warlike qualities that they supplied scarcely any recruits to the army; and they retained too much local patriotism to crowd into the official careers of senators or imperial servants. Although in the 1st century AD the astute Greek man of affairs and the *Graeculus esuriens* ("hungry little Greek") of Juvenal abounded in Rome, both these types were mainly derived from the population beyond the Aegean. The influx of Greek rhetoricians and professors into Italy during the 2nd and 3rd centuries AD was balanced by the large number of tourists who came to Greece.

The principal cities and sanctuaries received new buildings in the Greco-Roman style, magnificent work, though lacking in Classical distinction. For example, the emperor Hadrian completed the massive temple of Olympian Zeus at Athens begun six centuries earlier and built a library and an arch in that city. The Romans developed an enthusiasm for Greek sculpture and provided a market for great quantities of copies of Classical works (through which much of modern knowledge of ancient sculpture is known). In literature, the Greek language continued to develop, adding subtleties and distinctions in abstract thought as scholars wrote treatises on science, medicine, and philosophy. Noted Greek historians of the period included Polybius, who described the rise of Rome from 220 to 146 BC, and Plutarch, who began the art of biography with his comparisons of the lives of the noble Greeks and Romans.

In AD 15 the Greeks petitioned Tiberius to transfer the administration to an imperial legate. This new arrangement was sanctioned, but lasted only until AD 44, when Claudius restored the province to the Senate. The years 66 and 67 were marked by the long visit of the emperor Nero: in return for the flattering reception accorded to him, he bestowed freedom and exemption from tribute upon the country (this favour was speedily revoked by the emperor Vespasian). Important material benefits were conferred by Hadrian, who made a lengthy visit to Greece: besides erecting public works in many cities, he relieved Achaea of arrears of tribute and exempted it from various imposts; and he fostered national sentiment by establishing a new Panhellenic congress at Athens.

Visit by Nero

In the 3rd century Greece again experienced danger from foreign invasions. Already in 170 the Costoboci from beyond the Danube had penetrated into central Greece and sacked Eleusis, before being broken up by the local militia. In 253 a Gothic army unsuccessfully besieged Thessalonica. In 267 the province was overrun by Gothic bands in collusion with a Herulian navy; these troops captured Athens and advanced as far south as Sparta; but a new landing in the north was defeated by the emperor Gallienus, and the victories of the emperor Claudius II in 269 put an end to the immediate danger.

BYZANTINE HISTORY

Early Byzantine period. After the reorganization of the empire by Diocletian at the end of the 3rd century, Achaea occupied a prominent position in the diocese of Moesia. Under Constantine the Great, Macedonia was a diocese of the prefecture of Illyricum and was subdivided into the eparchies of Thessaly, Achaea (including some of the Ionian and Aegean islands), Epirus Vetus (including Corfu and Ithaca), and Crete, while the other Greek islands formed an eparchy of the diocese of Asia. A complex hierarchy of imperial officials was introduced, and the system of taxation elaborated, so as to yield a steady revenue. The levying of the land tax was imposed upon the *dekaprotoi* ("10 leading men") who, like the Latin *decuriones,* were entrusted thereafter with the administration in most cities. The tendency to reduce all constitutions to the Roman municipal pattern became prevalent.

Although the elevation of Constantinople to the rank of capital in 330 was prejudicial to Greece, which felt the competition of the new centre of culture and learning and had to part with numerous works of art to embellish it, the general level of prosperity in the 4th century was rising. Commercial stagnation was checked by a renewed expansion of trade resulting from the diversion of the trade routes to the east from Egypt to the Euxine

(Black) and Aegean seas. Agriculture remained depressed, and many small proprietors were reduced to serfdom; but the fiscal interests of the government called for the good treatment of this class, whose growth at the expense of the slaves was a step in the gradual equalization of the entire population under the central despotism. This prosperity received a sharp setback in a series of unusually severe earthquakes in 375.

Survival of pagan religion

The emperors of the 4th century attempted to stamp out by edict the old pagan religion, but, except for the decree of Theodosius I by which the Olympic Games were interdicted (394), these measures had no great effect and were not rigorously enforced. Paganism was long to survive in Greece—particularly in the Laconian Mountains. The sure footing gained by the Christian Church in Greece during the 4th century was strengthened by the judicious manner in which the clergy, unsupported by official patronage and often out of sympathy with the Arian emperors, identified itself with the interests of the people.

The schools of Athens in the 4th century still retained much of their prestige, and the Cappadocian fathers, for instance Basil of Caesarea, went there for their classical studies. In the 6th century Justinian forbade pagans to teach philosophy in Athens; and there appears to be no evidence that this work was taken over in Athens by Christian scholars, who seem at this time to have been attracted to the more flourishing schools of Gaza and Alexandria. But even if the schools of Athens declined in the early Middle Ages, there is ample evidence that the Byzantines prized their Hellenic tradition. Although they called themselves *Rhomaioi* (Romans) and were proud of their Christian orthodoxy, they never forgot their ancient Greek culture. Greece itself, however, was for a long time an obscure and neglected province, with no interests beyond its church and its commercial operations.

Barbarian invasions continued. The incursion of the Visigoths under Alaric (395–396) was accompanied by a systematic devastation that crippled Greece for decades. Vandal pirates raided the country in 466 and in 475. The Ostrogoths under Theodoric invaded Thessaly in 482. The Huns reached the Isthmus of Corinth in their invasion of 540 and Thermopylae in 558. The early part of the 7th century saw Greece invaded by Avars and Slavs, and numbers of Slavs settled in parts of Greece.

From the Isaurians to the Angeli, 717–1204. The emperors of the Isaurian dynasty in the 8th century continued the reorganization of the provinces into themes (administrative divisions). This reorganization appears to have been initiated by the 7th-century Heraclian emperors and subsequently extended to the rest of the empire. By the end of the 10th century, Greece was divided among the themes of Hellas, the Peloponnese, Nicopolis, Dyrrachium, Cephalonia, and Thessalonica, with the maritime themes of Sámos and of the Aegean Sea. (Crete had its own governor, but from the early part of the 9th century to 961 it was in Muslim hands.)

The Greek themes shared in the prosperity of the rest of the empire and had a flourishing economic life—particularly through the silk industry centred in Corinth and Thebes. In spite of imperial restrictions, the landed magnates (*e.g.*, the Sgouri of Nauplia and the Cantacuzini of Messenia) built up large family estates in Greece as elsewhere. Northern Greece was open to attack from the Bulgars late in the 9th and early in the 10th centuries, but by 1018 the Bulgarian Empire had been conquered by the Byzantines. Thessaly was occupied by nomad Vlachs at the end of the 11th century and became known as Great Walachia. Greece was also attacked from time to time by the ambitious Normans of southern Italy. Robert Guiscard took Durazzo (Dyrrachium) in 1082, whence his army advanced into Thessaly before it was driven out. During 1147–48 the army of Roger II of Sicily sacked Thebes and Corinth and took back to Palermo some of the Byzantine silk weavers from these flourishing centres of the industry.

Hostility between the civil and military powers in the 10th and 11th centuries was a major cause of the Byzantine Empire's weakness. Landowning provincial generals and entrenched bureaucrats in the capital put the defense of the frontiers second to their wars with each other. The system of themes, under which the military and civil authorities became subordinate to the general, had made the military class powerful. As generals became ambitious, they also became hungry for land. By the 10th century, these *strategoi* had acquired vast estates with large armed retinues; by the 11th century, they had almost eliminated the free peasantry. In retaliation against the generals who used troops against them in Constantinople, civil officials disbanded native levies and hired mercenaries who, however, proved unreliable.

The *pronoia* system

When, in 1081, the former general Alexius I Comnenus became emperor, he found it necessary to base his rule on the support of the powerful landowners. He used the *pronoia* system—granting revenue-producing properties in return for the performance of services—to establish the empire's military strength; but he also in effect entrusted the empire's armed forces to the control of the military aristocracy. Though the royal grant was not at first hereditary, it resembled a fief in the feudal system in most ways. *Pronoia* worked satisfactorily for about a century, but decentralization set in by the time of the Crusades. In the last days of the empire, the emperors made *pronoia* hereditary, often with exemptions from taxes and military service, to pacify the military aristocracy. Greeks under Frankish rule easily saw the resemblance between *pronoia* and fief, between archon and baron.

Late Byzantine period. With the crusaders' conquest of Constantinople in 1204 and the subsequent establishment of a Latin empire, Greece was split between Latin conquerors and Byzantine aspirants to the imperial throne. In the north the Greek ruler Michael Ducas Angelus Comnenus set up an independent kingdom consisting of Epirus, Aetolia, and Acarnania; the Epirotes for a time enlarged their territory and hoped to regain Constantinople. But when Constantinople fell to the rulers of Nicaea, Epirote territory and authority dwindled; and by 1340 what remained had either become part of the restored Byzantine Empire of the Palaeologi or fallen to the Angevins of Naples or to the Serbs. In the middle of the 15th century, after being overrun by Serbs and Italians, this territory was conquered by the Ottoman Turks. Thessalonica, Thessaly, central and southern Greece, and the islands were divided among the Latin conquerors after the Fourth Crusade, and a number of feudal principalities were set up. In the north the kingdom of Thessalonica was granted to Boniface of Montferrat, but it soon fell, first to Theodore, despot of Epirus, who took the title of emperor in 1223, then to the Byzantine emperor of Nicaea, John III Vatatzes, in 1246. After changing hands several times in the 14th and early 15th centuries, Thessalonica was captured by the Turks in 1430.

The Duchy of Athens (Attica and Boeotia) was granted as a fief by Boniface of Montferrat to the Burgundian Otto de La Roche, whose family kept it until 1308. It fell to the Almogávares in 1311. These Almogávares, otherwise known as the Grand Catalan Company, had entered Byzantine service as mercenaries on the conclusion of peace between the French and the Aragonese in Sicily (1302). They soon quarrelled with the Byzantines, however, and moved southward. Walter of Brienne, who had become duke of Athens on the extinction of the House of La Roche, engaged their services next, but quarrelled with them in turn, whereupon they defeated him in battle. Settling on his lands, they set up their own form of government in the duchies of Athens and Neopatras, with Aragonese princes from Sicily as their dukes. This Catalan-Aragonese domination lasted until 1388, when it gave way to the Florentine lord of Corinth, Nerio Acciajuoli, who in 1391 acknowledged Amadeo of Savoy as prince of Achaia (Achaea or Akhaïa) and called himself "lord of Corinth and the duchy of Athens and Neopatras." The Duchy of Athens was ruled by the Acciajuoli family (apart from a period of Venetian domination, 1395–1402), until it was conquered by the Ottomans in 1456–58.

Principality of Achaia

The most flourishing Frankish principality in Greece was that of Achaia, or the Morea (the Peloponnese), under the Villehardouin family. It was originally a fief of Boniface of Montferrat and was granted to William of Champlitte, but by 1209 it was held direct from the Latin emperor

by Geoffroy de Villehardouin with the title of prince. The Frankish rulers found the Byzantine system of *pronoia* not dissimilar from their own grants in fee and came to terms with the natives, who were allowed to hold their lands in return for military service to the conquerors. Toward the end of the 13th century, the splendid Frankish civilization of the Morea began to deteriorate. William of Villehardouin, prince of Achaia, had already been decisively defeated by Michael VIII (Palaeologus) at the Battle of Pelagonia (1259) and been forced after three years' captivity to cede the Moreot strongholds of Mistra, Old Maina, and Monemvasía (1262). The Palaeologi subsequently regained much more of the territory, which was reorganized in the 14th century as the Byzantine theme of the Peloponnese and governed by despots, usually members of the imperial family, with their centre at Mistra. In 1381, however, another band of mercenaries, the Navarrese Company, took possession of a considerable area in the west in the name of Jacques de Baux, a claimant to the Frankish principality; and from 1383, when Jacques died, the Navarrese themselves ruled there, until 1430, when they were overthrown by the Palaeologi. In the 15th century, however, there were renewed Turkish attacks, which the new Byzantine fortification, the Hexamilion across the Isthmus of Corinth, failed to stem, and Thomas and Demetrius Palaeologi were finally conquered by the sultan Mohammed II in 1458–60.

Venice long retained its Greek conquests. The partition of 1204 and subsequent acquisitions had given it a number of scattered possessions of strategic and economic value, ruled directly or by nominees or by special arrangement. The most important of these were: the Ionian Islands; Modon or Methóni (1206), Coron or Koróni (1206), Argos (1388), Nauplia or Návplion (1388), and Monemvasía (1464) in the Morea (Peloponnese); Crete (1204); and interests in the island of Euboea (shared with the Veronese family Dalle Carceri), as well as in numerous other islands and in the duchy of the archipelago. This duchy was organized under the leadership of the Venetian Marco Sanudo, who captured Náxos as early as 1204–05. By 1207 more than two dozen islands were held either by him or as fiefs from him, and henceforth the Latin archipelago was dominated by Venetian families such as the Sanudi and the Ghisi, the Barozzi, and the Crispi. Venice held Nauplia and Monemvasía until 1540, Crete until 1669 (two strongholds there until 1715), and the Ionian Islands until 1797.

The Frankish conquest brought the establishment of a Latin ecclesiastical hierarchy. The native Greek priests were, nevertheless, left undisturbed, and the higher Byzantine clerics continued to exist side by side with the Latins, though they often resided in Constantinople. It vigorously stimulated economic and cultural life in Greece and, on the whole, was accepted by the indigenous majority.

The political weaknesses of the Byzantine Empire brought a resurgence of Hellenism in Greece. The cities in Asia Minor were lost to the Turks, and Greece again became the centre of cultural life. Thessaloníki, renowned for classical studies, was famous for its schools and literary scholars. The churches of Athens were decorated in the Macedonian style by leading artists. Mistra, near the site of ancient Sparta, was the political and cultural centre of Byzantium.

Byzantine literary scholars rediscovered, after centuries of commentary and editing, an appreciation of the subject matter of their classical heritage. The most renowned intellectual of the age was George Gemistus Plethon of Mistra, a Neoplatonist philosopher who delivered a lecture, "On the Difference Between Aristotle and Plato," that inspired Cosimo de' Medici to found the Platonic Academy of Florence, which was to greatly influence the Italian Renaissance.

MODERN HISTORY

Greece under the Turks, 1453–1821. Turkish rule over Greece (*Tourkokratia*) is conventionally dated from 1453, when Constantinople was captured by the sultan Mohammed II and the last Byzantine emperor, Constantine XI, fell with his capital. In fact this date is no more than a symbol and a convenience, for the greater part of the territory of the empire was already in the hands of the Ottoman Turks, and some parts of the remainder were to be outside Ottoman control for generations and even centuries longer.

At Trebizond (Trabzon), on the Black Sea, a so-called Greek Empire survived until 1461; on the mainland, the despotate of Epirus outlived Constantinople in independence for a generation; and many of the islands fell to the Ottomans only in the following centuries—Rhodes in 1522, Cyprus in 1571, Crete in 1669, and Tēnos in 1715. The Ionian Islands almost entirely escaped Turkish occupation. But even those parts of Greek territory that escaped or delayed Turkish occupation were scarcely ever independent. The alternative was generally subjugation to the power of Venice, which proved generally more unpopular than that of the sultan. Venice gained control of the Peloponnese by the Treaty of Carlowitz in 1699, but the Turks were able to recover it within 20 years.

Ottoman conquest of Greek islands

Acquiescence in Turkish rule. The comparative popularity of Turkish rule during the first centuries of the occupation had a number of causes. In the first place, the rule of the early sultans was able and strong; at least it kept at bay the predatory powers of the west collectively known to the Greeks as Franks, who had shown, by diverting the ravages of the Fourth Crusade to Constantinople in 1203, that their fellow feeling for Christendom was barely skin-deep. In the second place, the doctrinal schism of Christendom had reached a degree of bitterness in the 15th century sufficient to make Muslims actually more welcome to Orthodox Greeks than the Catholics of Rome; and some held that the fall of Constantinople was no more than the just due of the Emperor and of the patriarch who had accepted reunion with Rome, on humiliating terms, at the Council of Florence in 1439. Mohammed II therefore astutely appointed Gennadius, the surviving leader of the party that had opposed reunion with Rome, to be patriarch under his rule.

In the third place, Ottoman rule was not initially oppressive in its incidence on the subject communities, at least in the golden age of the empire (whose apogee is generally identified with the reign of Suleiman the Magnificent, in the middle of the 16th century). The Greeks in particular enjoyed freedom in two interconnecting spheres, apart from the practice of their religion. One of these was freedom of trade, an occupation which the Turks thought inferior to the profession of arms and which enabled the Greek mercantile class to build up simultaneously contacts abroad, wealth, a tradition of independence, and a merchant fleet that could be readily converted to warlike purposes. The second great freedom was that of education in the Greek language, which went hand in hand with the freedom of the church.

Another reason the Greeks acquiesced so readily for so many generations in the Turkish occupation was the nature of the Ottoman political system, which, paradoxically, led to their later struggle for independence. The system was based on the traditions of the Turkish tribes' nomadic past, and although it adapted to more settled conditions, it never fully assimilated the conceptions of fixed geographical frontiers or the nation-state. The one constant factor was the centralization of all power in the hands of the sultan. His empire was divided for administrative purposes into provinces governed by his appointees; Greece was distributed among six governors. All the land remained the property of the sultan, who could dispose of it as he wished in fiefs allotted to Turkish settlers (usually loyal military adherents) or permit his non-Muslim subjects to live on it on payment of a land tax. There grew up piecemeal exceptions to the system. Some important garrison towns were subordinated directly to Constantinople instead of to the provincial governor; some areas achieved a special status (Chios, for instance, becoming a private appanage of the empress-mother of the sultan); and other areas, because of their inaccessibility, were left virtually to their own devices (*e.g.*, parts of Crete and of the Pindus Mountains and some of the smaller islands). A varying measure of local autonomy in the administration of justice was also tolerated, especially in the Peloponnese.

Taxation under Ottoman rule

At the height of its power, however, Ottoman rule was arbitrary and unpredictable, being too dependent on the character of individuals, especially on that of the sultan. The Greeks' principal obligations were to pay the capitation tax (*kharaj*), which simply entitled them to remain alive in the sultan's dominions, the land tax, the tax on commerce, and occasional special levies according to their circumstances and way of life; and to contribute male children (about one out of five on average) to the sultan's private service, to be brought up as Janissaries. The last was a cruel practice (though not always to the worldly disadvantage of the Greek children, who might rise to great heights), and it was abolished before the end of the 17th century. The vice of the system was that the Greeks received no rights—not even the elementary rights of justice and security—in return. The result was, on the one hand, that they turned increasingly to their own resources to manage their affairs (especially to the priests and to the leaders of their own communities, known as *prokritoi* or primates); and, on the other, that in many areas lawlessness became endemic, in the form of the *klephtai*—whom Greek tradition regards as Robin Hood-like outlaws of the mountains—and the scarcely more law-abiding *armatoloi* ("men at arms" or "gendarmes"), whom the Turkish government indiscriminately licensed to put down the outlaws. Finally, the Greeks had one other course open to them: this was to penetrate and to usurp the system themselves by climbing the rungs of power that the corps of Janissaries and other opportunities opened to them. By the 18th century they had done this so successfully that four of the great offices of state (dragoman of the Porte, dragoman of the fleet, and voivode, or governor of each of the two Transdanubian provinces, Moldavia and Walachia) had become virtually Greek preserves. The Greeks who infiltrated the administration system in this way formed a distinct colony in Constantinople, known as the Phanariotes (from the Phanar, the quarter in which they lived).

Such manifestations of opposition to Turkish rule as occurred up to the 18th century were as spasmodic and unpurposeful as Turkish oppression. Greek uprisings were generally in response to deliberate and unscrupulous provocation by external powers in the course of great wars fought for no interest of the Greeks, who were then generally the victims of both sides. At the naval battle of Lepanto in 1571, for instance, between the Turks and the forces of Don John of Austria, Greek islanders were by force of circumstances serving in both fleets; and mainland Greeks who rose in revolt at the instigation of Don John were put down without mercy by the Turks. In the following century the invasion of the Peloponnese by the Venetian general Francesco Morosini (whose guns were responsible for the destruction of the Parthenon at Athens in 1687) aroused not a spark of favour among the Greeks; and the return of the Turks to the Peloponnese in the following generation was accepted as the lesser evil.

It was not until the intervention of the Russians in the 18th century that a serious revolt was promoted on Greek territory; and again the motives were unconnected with any real demand for national independence. Meanwhile, the influential Greeks of Constantinople were in effect pursuing quite a different policy of their own, which was not to disrupt the Ottoman Empire but to take it over as a going concern.

Rising of 1770 and beginnings of nationalism. The rising in the Peloponnese in 1770 is virtually the beginning of modern Greek history. It was precipitated by the empress Catherine the Great of Russia as a diversion in one of the recurrent Russo-Turkish wars; but for the first time it struck a spontaneous response from the Greeks (especially the Peloponnesians) and thus came near to success.

Russian interest in Greek affairs

The Russian right to interfere in Greek affairs rested ostensibly on the possession of a common religion, which the tsars regarded themselves as entitled to protect, and on the inheritance of the Byzantine tradition as a result of the 15th-century marriage between the niece of the last emperor of Constantinople and the grand prince Ivan III of Muscovy, the first Russian ruler to call himself tsar. In practice, the real motive lay in the advantage that a Russian liberation (which meant annexation) of Greece would give the tsars over the Ottoman rulers, by opening a back door to Constantinople.

The Greeks, however, welcomed the approaches of the tsar's agents for reasons of their own. In the Peloponnese, society was relatively advanced and ripe for self-rule and, being relatively well populated, the area suffered more than most from the land hunger that resulted from the rule that no land could pass from Turkish into Greek possession. The combination of a politically conscious and aggrieved upper class and a landless and therefore lawless lower class led naturally to a revolutionary mood.

The revolt of 1770, prepared and incited by Russian agents, took the Turks by surprise and was at first successful. But Russian help was late and inadequate (having to come by sea around Europe and through the Mediterranean); and, with the help of a bloodthirsty army of Albanians, the Turks eventually suppressed the rising in the Peloponnese, the only area in which it reached serious proportions. The Russians quickly abandoned the Greeks in order to pursue the war with Turkey elsewhere and brought it to a relatively successful conclusion in the Treaty of Kuchuk Kainarji (1774). An ambiguous clause in this treaty gave the Russians a right "to make representations" on behalf of the sultan's Christian subjects at Constantinople, and this provided the excuse for all future Russian interventions in Greek affairs.

After about nine years of barbarous repression by the Albanian forces of the sultan, the Peloponnesians recovered a normal level of existence and prosperity. But when Catherine the Great tried again in 1786 to provoke a similar rising against the Turks, the Peloponnesians had learned their lesson too well to respond. The only response came from elsewhere. The inhabitants of Suli (Souli), Epirus, then in southern Albania, rose against the sultan and were savagely suppressed by the provincial governor, Ali Pasha, from his capital at Ioánnina. The Russians again abandoned their allies, but Ali Pasha's appetite was whetted by success, and within the next 30 years he proceeded to carve out for himself a large semi-independent slice of the sultan's empire. This remarkable figure unwittingly played a large part in making possible the Greek revolution of 1821.

The climate for revolution

Between the risings of 1770 and 1821, history underwent a great change that decisively shifted the balance in favour of the Greeks. The American and French revolutions, the rise and fall of Napoleon, the conflict of the great powers in the Middle East, the outbreak of nationalist revolutions as far away as South America and as near as Serbia, the weakness of the Ottoman Empire as revealed by the virtual secession of Ali Pasha, the incitements of European liberals and poets, not only foreign (such as Lord Byron) but also native (such as Rhigas Pheraios), and the awakening of a national consciousness by native scholars (such as Adamantios Korais) as well as by merchants, politicians, and priests all contributed to the outburst of 1821. The one factor that played relatively little part in bringing about the revolution was immediate Turkish oppression. It was rather the enfeeblement and relaxation of tyranny that fed the appetite for liberty, as it was the apparently imminent dissolution of the Ottoman Empire that offered the opportunity of escape.

But the positive motives at work among the Greeks were dangerously confused. The aspirations of the different classes and regions were unreconciled. The Greek aristocrats of Constantinople, of the Transdanubian provinces, and of the tsar's court had little in common with the peasants, priests, and brigands of central Greece or with the wealthy merchant families of the islands; and no reconciliation had been attempted, or was possible, between those who looked to Russia, to England, to France, or even elsewhere for patrons of Greek independence.

All the great powers had a more or less definite interest in Greek affairs by 1821. In Russia there were considerable Greek colonies, including that at Odessa, which in 1814 founded the Philikí Etaireía (Friendly Company) with the express object of promoting a Greek revolution; and there were also highly placed Greek officials, including the Emperor's foreign secretary Count Ioánnis Kapodístrias

and his aide-de-camp Prince Alexandros Ypsilantis. The Austrian Empire was hoping to encroach eastward and southward as the Ottoman Empire disintegrated and was seeking to play the same role in the northern Balkans that Russia sought to play in Greece. England and France alternated in activities throughout the Middle East during the Napoleonic Wars, wooing Ali Pasha, the sultan, and the dissident Greek leaders with arms, money, and promises by turns. The fate of the Ionian Islands was typical of the whole complex of forces at work. They were taken by the French from Venice in 1797 and then occupied in turn by the Russians (1799), the Turks (1800), the French again (1807), and the British (1814) in the space of 20 years; and one of their dependencies, the mainland port of Parga, was even ceded to Ali Pasha by the British in 1817. From all such manoeuvres of power politics, the Greeks were simply the sufferers; but experience was teaching them how to take advantage of these rivalries. The Ionian Islands, under British control, were havens where armed irregulars from the mainland could take refuge under pressure. Russia provided a base from which the Philikí Etaireía could mature its plans, as well as a school of statecraft where Greeks could learn their political trade. Vienna was another haven of refuge and planning for Greek émigrés. The French were still reckoned the greatest soldiers of Europe even after Waterloo; and the fall of Napoleon released many military adventurers from whom the Greeks (as well as the Turks and Egyptians) could learn the art of war. Added to this was the newborn European liberal prejudice toward freedom, which provided the Greeks with money, arms, propaganda, and philhellenes to fight at their side. Only the opportunity of revolt was needed, and this was provided by the rebellion of Ali Pasha against the sultan Mahmud II in 1820. While the Turks were preoccupied with this threat, the Greeks in their turn rose to war in March 1821.

The Revolution of 1821 and its results. On March 25, 1821, Bishop Germanos raised the banner of revolt at the monastery of Aghia Lavra in the Peloponnese. The ensuing revolution went through three phases: local successes in 1821–25, the crisis caused by the intervention of Muḥammad ʿAlī of Egypt in 1826–28, and a period of European intervention ending in Turkish recognition of Greek independence in 1832.

The revolution, 1821–29. At the beginning the revolution spread in a flash, embracing the Peloponnese, central Greece, and the island of Crete. By the summer of 1822 the revolutionaries had captured Missolonghi, Athens, and Thebes. At that point the revolution was checked by the Ottomans and by internal conflicts among the Greeks. The Ottomans hanged the patriarch Gregorios in Constantinople and suppressed the uprising in Thessaly, Macedonia, and Mt. Athos. A stalemate prevailed until the beginning of 1825. There were dissensions among the Greek insurgents, and the aspirations of different classes and regions were too various to be reconciled. In 1822 two Greek governments existed: one on the mainland dominated by Theodóros Kolokotrónis, and the other at Ídhra led by Geórgios Kountouriótis and Aléxandros Mavrokordátos. By 1824 open civil war prevailed in Greece. A turning point came with the intervention of Egyptian forces under Muḥammad ʿAlī's son Ibrahim Pasha, who landed on the Peloponnese with a large army in February 1825. By June 1827, burning and plundering, he seemed about to reconquer Greece. Missolonghi, the key fortress of Corinth, fell in April 1826 after a heroic defense, marked by the death of the English poet Lord Byron.

The Greek revolution was rescued by the powers of Europe, who, motivated by complex political factors, offered mediation to the belligerents in 1827.

Although the Battle of Navarino (October 20, 1827), fought to enforce the Treaty of London (July 6, 1827), made the independence of Greece a certainty, another two years passed before the fighting ended, and nearly five before the new state took shape. Navarino was not the end of a war among the great powers but in effect the beginning of a new one. The Sultan proclaimed a holy war (*jihad*) in December 1827, and Russia declared war on Turkey on April 26, 1828. France and Great Britain reluctantly withdrew their ambassadors from Constantinople. It was not until the Russian armies had almost reached Constantinople that, by the Treaty of Adrianople (September 14, 1829), the Sultan formally accepted the independence of Greece as prescribed by the Treaty of London. From this date all the great powers and other contestants finally resigned themselves to the two complementary facts that Greek independence could not be reversed and that it did not entail the general dissolution of the Ottoman Empire.

First phase of independence, 1829–64. Neither the boundaries nor the constitution of the new Greek state was yet settled. During the Russo-Turkish War of 1828–29, desultory fighting continued on the Greek mainland, almost neglected by the preoccupied powers, and even the army of Ibrahim Pasha remained at large in the Peloponnese until a French expeditionary force was sent there in 1828 to ensure its withdrawal. While the conference of London debated at its leisure the borders of the new state, the Greeks set about enlarging their territory and lobbying to secure the addition of important islands such as Sámos and Crete. They were helped by the military skill of British free lances, such as Sir Richard Church and Frank Abney Hastings, as well as by the sympathetic connivance of French regulars, to install themselves well to the north of the Gulf and Isthmus of Corinth on both the western and eastern flanks of the Greek mainland during 1828–29, though they could not recover strongly held fortresses such as Athens. Foreign-born Greek leaders, such as Aléxandros Mavrokordátos and Dimitrios Ypsilantis (the brother of Alexandros Ypsilantis), also now played a notable part in consolidating Greek expansion, as well as popularizing Greek ambitions in terms familiar to the European courts; and, above all, Kapodístrias, who had been elected provisional president of Greece by the Third National Assembly (April 1827) and arrived to assume office in January 1828, and who conducted Greece's foreign affairs with the utmost skill and devotion.

Kapodístrias and his successors. Kapodístrias had a clear, autocratic idea of the way to run his new country, with a strong, centralized government that suppressed the last traces of local self-government; and, without actually suggesting that his position should be made permanent, he regarded himself as irreplaceable. Meanwhile the great powers, represented by the Conference of London, sought a more conventional settlement.

At the time of Navarino, they still had in mind a ruler of Greece who would remain the sultan's vassal. The Treaty of Adrianople, by confirming Greece's unqualified independence, made this impossible, and on February 3, 1830, the throne of Greece was offered to Prince Leopold of Saxe-Coburg (the future king of the Belgians, uncle of the future queen Victoria), who at first accepted. Kapodístrias, however, subtly hinted to him that the terms of the offer were inadequate to ensure Greece's viability, and Leopold withdrew his acceptance. Kapodístrias continued his presidency while another candidate was sought, and he grew steadily more unpopular, not with the ordinary people, who benefitted by his firm rule, but with the "primate" families (such as Kountouriótis of Hydra and Mavromikhalis of Maina), who had expected to inherit the Turkish feudal system under their own mastership; with the liberal-minded constitutionalists (led by Mavrokordátos, who refused to take office under the president); and with most of the irregular leaders of the war. To the natural dissensions of the Greeks were added the activities of the three official residents, Russian, British, and French.

Opposition of the "primate" families

Kapodístrias was assassinated at Nauplia (the first capital of Greece) on October 9, 1831, by two members of the Mavromikhalis clan. He was succeeded by a triumvirate consisting of his brother Avgoustinos and two of the wartime leaders, Ioánnis Kolettis and Kolokotrónis, but anarchy soon supervened. Avgoustinos fled the country in April 1832, and Kolettis and Kolokotrónis each tried to establish a government of his own (the former endorsed by Mavrokordátos, and each supported by a private army of irregulars). The chaos was mitigated only by the continued presence of a reduced French force in the Peloponnese, until at last the Conference of London reached its final decision, which was embodied in a convention signed in

May 1832. By this convention Greece was established as an independent kingdom, under the protection of Great Britain, France, and Russia, with its northern boundary along the Árta-Vólos line, and including some of the Aegean Islands (but not Crete or Sámos, for which Kapodístrias had fought in vain). Louis I of Bavaria, who was also a party to the convention, accepted the Greek throne on behalf of his 17-year-old son Otto, who arrived at Nauplia on February 6, 1833, and endeared himself by at once adopting the Greek national costume and the Greek spelling of his name, Othon.

Otho. The accession of King Otho (as he is usually called in English), backed by a generous loan from the protecting powers, was looked upon as the beginning of a new golden age for Greece. There were certainly agreeable episodes to record in the immediately following years: the transfer of the capital from Nauplia to Athens (1834); Otho's coming of age (1835) and marriage to Amalia of Oldenburg (1836); the founding of the University of Athens (1837); and the establishment of diplomatic relations with Turkey (1839).

But there were constant disorders in different parts of the kingdom, chiefly the result of the discontent of wartime leaders who had been set aside in favour of Otho's new Bavarian administrators. There were major outbreaks of pillage on the mainland in 1834–36 and again in 1839, and in the latter year conscription was first instituted. There was a revolt in Crete in 1841, a year after the island had been restored from the Egyptian province of Muḥammad 'Alī to the direct rule of the sultan Abdul-Mejid; and although this was legally no business of the new Greek kingdom, it was one of many incidents that reminded the Greeks that only a small part of their racial and religious kin was independent.

Above all, there was resentment of the alien and untactful rule of the Bavarian clique led by Graf Josef Ludwig von Armansperg, who was in effect absolute ruler of Greece from 1834 to 1837. It was not until December 20, 1837, that a Greek, Konstantinos Zographos, was first appointed as the king's chief minister, and even thereafter the Bavarians continued to hold the Ministry of War and to penetrate the administration effectively at many levels.

Revolution of 1843

Resentment came to a head in 1843. On the night of September 14–15 the royal palace was surrounded by a rebel force led by veterans of the War of Independence, whose chief demands were the removal of the Bavarians and the establishment of a constitution. It was a typically Greek revolution, conducted almost without bloodshed, with almost no bitterness (except against the Bavarians), and with a strong streak of conservatism and personal affection for the King. As soon as Otho had given way, he became a national hero, and the occasion a national holiday. A national assembly was convened (attended significantly by representatives from the unliberated areas of Thessaly, Macedonia, and Epirus) that drew up a constitution providing for a lower house (*vouli*) and a senate (*yerousia*). This constitution, to which the King took his oath on March 16, 1844, remained in force for 20 years.

There followed a period of domestic quiescence, marred only by disagreeable episodes in Greece's foreign relations. With Great Britain there were a number of points of friction that did little credit to the protecting power. The British government was entitled to claim, as it vigorously did, the payment of interest on an international loan in 1847, as well as compensation for losses arbitrarily inflicted on British subjects including the philhellene historian George Finlay and the Gibraltar Jew Don Pacifico; but it was not entitled to enforce the claims, as Lord Palmerston insisted on doing in January 1850, by ordering the Mediterranean fleet to blockade Piraeus and openly hinting an intention to annex two minor islets alleged to have formed part of the Ionian Islands. Palmerston's undignified conduct nearly brought down the government; the British claims were reduced and met, but Otho came out of the episode with greatly enhanced prestige for his courageous refusal to be bullied. Unfortunately this success led him into a more foolhardy act of defiance four years later and to a passionate devotion to the cause of Greek irredentism known as the Megali Idea (Great Idea).

In the Crimean War the British and French sided with Turkey against Russia. Imprudently, but for obvious reasons, the Greeks sided with Russia, and Otho took the lead with strong popular support in an attempt to enlarge Greece's territory. There had been no major breach of the peace between Greece and Turkey for 25 years, though pillage and frontier incidents had led to a brief rupture of diplomatic relations in 1847. Now the Greeks seriously attempted to conquer Thessaly and Epirus. In January 1854 they entered Epirus and defeated the Turks at Péta; in Thessaly they had little success. The Turks sent an ultimatum to Greece in March and began to expel Greeks from Smyrna and Constantinople in reprisal. In May British and French troops landed at Piraeus to enforce Greek neutrality, and the occupation lasted until 1857. Greece's frontiers thus remained unchanged and its appetite unappeased.

At the same time Greece's relations with Great Britain were exacerbated by the question of the Ionian Islands, which had been left under British protection by the treaty of November 5, 1815. The national feeling of the Greek inhabitants of the islands had reached a dangerous intensity, which produced a serious rising on Cephalonia in 1849. Their future now became fortuitously connected with the unhappy end of Otho's reign. Otho misused his popularity to behave autocratically at home, while distracting his subjects with the hopes of acquiring territory abroad. Despite a succession of patriotic Greek prime ministers, including several heroes of the War of Independence (Mavrokordátos, Kolettis, Kitsos Tzavelas, Geórgios Kountouriótis, Konstantínos Kanáris), the King never fully accepted the role of a constitutional monarch. Opposition to him grew, first breaking out in a mutiny in the garrison of Nauplia on February 13, 1862.

This was suppressed, but by October trouble was again brewing. While Otho and Queen Amalia were on a tour of the Peloponnese, the revolt began in Aetolia-Acarnania on October 19 and spread to Athens on October 22. Otho arrived back in his naval frigate on the following day to find in power a provisional government, which had proclaimed the deposition of Dimitrios Voulgaris and also its intention to maintain the monarchy. Otho never set foot in his capital again. The protecting powers recognized that the Greeks were acting within their rights in deposing their monarch and searched for a successor. More than a year passed before the right man was chosen.

Otho's loss of power

Various candidates were ruled out. Finally the search ended with the choice of the 17-year-old William, second son of the heir to the throne of Denmark. On June 5, 1863, representatives of the three powers and Denmark signed a protocol in London recognizing his election (which the Greek assembly had enthusiastically carried on March 30) to the vacant throne, on the explicit understanding that the Ionian Islands would be ceded to Greece.

Treaties of the protecting powers

The treaty of cession was signed in London on November 14, 1863, by representatives of Great Britain, France, Austria, Prussia, and Russia (the powers that had established the protectorate) and confirmed by a further treaty between Greece and the three protecting powers (Great Britain, France, and Russia) on March 29, 1864. The new king, who took the title "George I of the Hellenes" (his predecessor had been "king of Hellas"), had arrived in Athens on October 30, 1863, and the British evacuation of the Ionian Islands was completed by the following June.

The new monarchy, 1863–1924. In the half century of King George I's reign (1863–1913), Greece grew to the status of a power in its own right (though the "protectorate" of Great Britain, France, and Russia nominally lasted until 1919). The period was notable for the most considerable increases of Greek territory. Thessaly, most of Greek-inhabited Epirus and Macedonia, Crete, and the majority of the Aegean Islands (except the Dodecanese) were added to the kingdom of the Hellenes. These advances were not achieved without setbacks, which were chiefly due to political instability at home; future troubles were foretold by the constitutional dispute between supporters and opponents of the monarchy and the mutual hostility between Greece and other successor states of the Ottoman Empire (chiefly Bulgaria and Albania). It was

also in this period, nevertheless, that Greece produced its first two statesmen of recognized European eminence: Kharílaos Trikoúpis (1832–96) and Eleuthérios Venizélos (1864–1936).

On his accession, the new king inherited a legacy of political confusion bordering on anarchy and even violence. A succession of ministries came and went during his first year. But a clean break with the past came with the inauguration of the new constitution, to which he took his oath in November 1864. It was highly democratic. It abolished the senate, replacing it with a crown-appointed council of state (abolished in its turn a year later); it established elective local government; and it defined the position of the king, which was in effect that of a passive instrument of the will of the people.

This system came to be described, by an inversion of the usual term, as a *vasilevomeni dimokratia* ("monarchical democracy"); and there is a happy irony in the phrase, since *dimokratia* is also the Greek for "republic." King George I operated this constitution conscientiously and successfully for 47 years, until it was revised in 1911. He was also luckier than his predecessor in not being accompanied by a foreign garrison; and although he had Danish advisers in his train, the most unpopular of them, Count Wilhelm Carl Sponneck, left Athens in 1865. Greece was at last truly free.

The Cretan rising, 1866. Though relative tranquillity was restored at home, the new reign's troubles abroad began soon. In 1866 the Cretans rose in revolt against Turkish rule, which in its decline had become truly oppressive. The sultan, Abdülaziz, sought not only to suppress the revolt but to transfer Crete back to Egyptian rule. A self-constituted Cretan assembly proclaimed its union with Greece, and although the King and government could not risk accepting the decision, many volunteers from the mainland went to support the islanders in arms.

The Cretans appealed to the great powers, of whom Great Britain alone was opposed to the solution of a plebiscite. In particular, European opinion was stirred by the heroic end of the monastery of Arkadi, whose abbot blew it up with the garrison and its attackers together rather than surrender alive. Though the revolt was crushed, the intervention of the powers obliged the sultan to grant administrative reforms, which were embodied in the organic statute of 1868.

The Cretan problem, which coincided with George I's reign, was thus shelved again. The sultan took his revenge on the Greeks two years later by creating in 1870 the Bulgarian exarchate, a separate ecclesiastical organization for Bulgaria, independent of the Greek Orthodox patriarchate that had been for so many centuries the unifying force of the Ottoman Empire's Christian subjects.

The Russo-Turkish War, 1877–78. The second crisis of the new reign was precipitated by the Russo-Turkish War of 1877–78. As in the Crimean War, many Greeks saw in it an opportunity to acquire Thessaly, if not Epirus and Crete as well. But a coalition government, in which Trikoúpis was foreign secretary, held them in restraint, in the expectation that the great powers would sympathetically consider Greek claims at the peace settlement. Russian victories swept aside these restraints. The coalition government fell, and its successor was only withheld from armed intervention by British mediation and by the collapse of Turkish resistance in January 1878. In the Treaty of San Stefano (March 3, 1878) between Russia and Turkey, Greece's territorial claims were ignored in the interests of creating what was called Greater Bulgaria as a Russian satellite, with access to the Mediterranean through the Aegean Sea and practically the whole of Macedonia incorporated in it.

The other powers refused to accept this settlement, which was defeated by the Congress of Berlin (June 1878); but still Greece received no satisfaction beyond an injunction to the sultan Abdülhamid II to reach an agreement with King George for the modification of their common frontier. The fulfillment of this injunction was dragged out until 1881, when a conference at Constantinople finally defined a new frontier between Greece and Turkey from the southern slopes of Mt. Olympus in the east to the Árakhthos River in the west, giving Greece the greater part of Thessaly and a small corner of Epirus.

Trikoúpis. Between the acquisition of Thessaly in 1881 and the next great foreign crisis in 1896, the domestic scene in Greece was dominated by Trikoúpis. He became prime minister early in the 1880s and remained in power, with brief intervals, until his resignation in January 1895. During this period great advances were made in Greece's domestic administration. Roads and railways were extended, the merchant navy was expanded, the Corinth Canal between the Peloponnese and mainland Greece was completed, pillage was virtually suppressed, and the financial administration was put on a sound basis that attracted the confidence of foreign investors. These investors, however, brought with them the intervention into Greek affairs of their stronger European governments whenever the protection of those investments was deemed necessary.

Progress was subject to abrupt setbacks whenever Trikoúpis' chief rival, Theodoros Diliyiannis, a crafty but hotheaded parliamentarian, succeeded in replacing him. Trikoúpis, upon his return to power, dealt competently with the setbacks caused by Diliyiannis' militant policy and retained the confidence of the King and people. Even after Trikoúpis' party had been defeated in a general election in 1890, the King reappointed him prime minister in 1892, after dismissing Diliyiannis for incompetence.

Rivalry of Theodorus Diliyiannis

Crete and Macedonia, 1896–1912. It was after Trikoúpis' death (in April 1896), when Diliyiannis was again prime minister, that the next major crisis in Greece's foreign affairs occurred, again over Crete. The chronic bitterness between Cretans and Turks broke out in violence and bloodshed at Khaniá (Canea), the administrative capital, in May 1896.

This time it was Diliyiannis' turn to exercise restraint and to accept Turkish promises of reform. But these were not carried out, and the deteriorating situation culminated in a further outbreak of violence in February 1897. Diliyiannis announced that Greece could no longer remain indifferent and sent an armed force to annex the island. Thereupon Great Britain, France, Russia, Italy, Germany, and Austria proclaimed an international protectorate and landed troops. The intervention of the great powers exasperated Greek public opinion to the point of demanding war on Turkey; and in March 1897, while the expedition to Crete was still indecisive, the crown prince Constantine put himself at the head of the Greek forces in Thessaly in readiness for war. Turkey declared war on April 17, and Greece was defeated in a brief campaign. Peace was signed on December 4, 1897. Greece naturally had to withdraw from Crete, which remained under the occupation and protection of the great powers, and minor modifications were made to the northern frontier in Turkey's interest, but the indemnity originally demanded was reduced from £T10,000,000 to £T4,000,000.

The repulse of Greek claims to Crete and Macedonia was looked upon as no more than temporary, and the two territories largely occupied political attention in Greece for the next 15 years, until the chance of a settlement came. In Crete the disturbances of 1896–97 represented a slight immediate advance for the Greek cause. The protecting powers (whose number was reduced to four by the withdrawal of Germany and Austria) proclaimed the administrative autonomy of the island; and after a further outbreak of violence in September 1898 (in which the British vice-consul at Iráklion and a number of British soldiers were killed), they insisted that all Turkish troops should leave the island and invited the King's second son, George, to become high commissioner under the suzerainty of the sultan. Prince George arrived in Crete in December 1898, for a term of three years prolonged until 1906. His tenure was marred and, finally, terminated by a dispute with a group of Cretan politicians, led by Venizélos, who seceded in March 1905 to set up a "provisional government" at Thérisson.

Prince George was replaced in 1906 by Alexandros Zaimis, a former prime minister and one of the most respected statesmen of modern Greece, under whose auspices Venizélos became the leading figure in the Cretan administration. Tranquillity and order were so effectively

restored that in July 1908 the great powers felt justified in beginning the withdrawal of their forces.

Effects of the Young Turk Revolution

At almost the same date, an event took place at the other end of the Greek world that upset the powers' calculations. This was the Young Turk Revolution, which set up the Committee of Union and Progress at Salonika, the principal town in Macedonia. Having compelled the sultan Abdülhamid to restore the lapsed constitution of Turkey, the committee then deposed him in 1909 and ruled the Ottoman Empire in the name of his successor, Mohammed V; but toward the Christian minorities it pursued a policy that resurrected all the Greeks' latent fears for the future of Macedonia. Since the end of the 19th century, a state of undeclared civil war among Greek, Bulgarian, and Turkish irregulars had prevailed in Macedonia. The Young Turk Revolution worked a widespread transformation in the whole Balkan scene. It brought together Greeks, Bulgars, and Serbs in the common interest of frustrating the "Ottomanization" of Macedonia; it provoked the Cretans yet again to declare the union of Crete with Greece in October 1908 and to send a delegation (including Venizélos) to Athens to give effect to their decision; and it led to the formation in the Greek Army of the Military League (May 1909) in imitation of the Committee of Union and Progress. Although the Greek government pursued the correct course of disowning the Cretan proclamation of union, the Military League succeeded in forcing many of its wishes on the King, including the summoning of Venizélos to Athens in January 1910.

From March 1910, when the Military League dissolved itself on the completion of its task, Greece moved swiftly to the climax of the Balkan Wars. A national assembly was elected in August, at Venizélos' instigation, to revise the constitution, but when it proved unamenable, Venizélos, having become prime minister for the first time on October 18, dissolved the assembly and held a second election in December.

From this election, which gave him a large majority, dated Venizélos' extraordinary ascendancy over the Greek political scene. His revised constitution was promulgated in June 1911, his peace was made with the royal family, the election of a Chamber of Deputies under the constitution of 1911 gave Venizélos a majority of five to one in March 1912, and he was ready to set about the creation of the Balkan League, which was to destroy the Ottoman Empire in Europe. Still, he would not admit the Cretan representatives to the new chamber until he was ready and strong enough to face the consequences.

Balkan wars, 1912–13. In the spring of 1912 Bulgaria concluded treaties with Serbia and Greece; Serbia and Greece had no conflicting interests and no formal alliance. In October the three governments presented joint demands to Turkey for reforms in Macedonia, which were unacceptable. Turkey declared war on Bulgaria and Serbia on October 17. Venizélos admitted the Cretan representatives to the chamber and declared war on Turkey on October 18. The First Balkan War lasted seven months, interrupted by an armistice (December–January) for peace negotiations in London, in which Greece participated without accepting the armistice. The war was a success for the Balkan allies, particularly for Greece, whose armies captured Salonika, Préveza, Párga, and Ioánnina, besides the acquisition of Crete, Sámos, and other islands. Constantine succeeded to the throne on March 18, 1913, when King George I was assassinated.

Succession of Constantine

Salonika was the point of friction among the allies, of whom the Bulgars at least also hoped to acquire it. Even before the defeat of the Turks was complete, Greece and Serbia began to negotiate a treaty of alliance against the danger of attack by their ally, Bulgaria. The First Balkan War was terminated on May 30, 1913, by the Treaty of London, which abolished the Turkish Empire in Europe. It ceded Macedonia to the Balkan League as a whole and left the disposition of the Aegean Islands to the powers. Two days later, the Greco-Serbian alliance was concluded; and within a month the Second Balkan War had begun with a Bulgarian attack on the two allies.

The Second Balkan War, in which Romania joined Greece and Serbia, ended in a defeat of the Bulgars and a considerable recovery of lost ground by the Turks. It was terminated by the Peace of Bucharest on August 10, 1913. Bulgaria retained a small part of Macedonia and, at Dédéagatch (Alexandroúpolis), an outlet to the Aegean Sea in western Thrace. Relations between Greece and Turkey remained bad, and the Balkans seemed to be again on the brink of another convulsion when the outbreak of World War I in August 1914 merged their parochial problems with the great upheaval of Europe.

World War I and Asia Minor. World War I opened up a bitter difference between the King and Venizélos that divided Greeks for a generation into royalists and republicans. The two men disagreed on the likely course of the war and consequently on the interpretation to be put on the Greco-Serbian treaty in the event of Bulgaria's joining the Central Powers in hostilities against Serbia. Venizélos hoped for a victory of the entente. Constantine was more uncertain of the outcome and therefore preferred neutrality, with a sympathy toward the Central Powers. The general staff, especially the chief of staff, Col. Ioannis Metaxas, aided the King with their military judgment. Though the crisis was not reached until September 1915, when Bulgaria mobilized to attack Serbia, the two attitudes were clear from the start. On August 7, 1914, the King assured the German emperor William II (whose sister Sophia he had married) that he, though officially neutral, was on Germany's side in sympathy; and two weeks later Venizélos indicated to the entente that his sympathies were with them and that Greece would enter the war on their side if Turkey did so against them. Turkey in fact entered the war against the entente in November 1914.

Royalist and republican rivalry

At this juncture the King's will prevailed, supported as he was by his Chief of Staff, and he was able both to keep Greece neutral and to retain Venizélos as his prime minister. The struggle between Venizélos and the general staff came to a head in March 1915, when the King dismissed his prime minister rather than accept the resignation of Metaxas. In Venizélos' place the King nominated Dimitrios Gounaris, who unfortunately chose to represent Venizélos as motivated by spite against the royal family. Gounaris retained power even after a general election in June had given Venizélos' party a small majority. Venizélos was recalled to office only in August and dismissed again in October, but in the interval there had occurred the decisive events of Bulgaria's mobilization to attack Serbia, Greece's countermobilization, and the arrival at Salonika of Allied forces on Venizélos' invitation. To restore the situation the King appointed Zaimis prime minister, with instructions to repudiate Greece's obligations under the Greco-Serbian treaty. Even Great Britain's offer to cede Cyprus to Greece did not sway the new government, and it was soon withdrawn. Greece remained neutral, the King dissolved the Chamber, and the Liberals abstained from the elections of December 1915 in protest.

Neither Venizélos nor the Allies took further drastic measures against the King's government until after the surrender of Ft. Rupel, on the Bulgarian frontier, to the Germans and Bulgars in May 1916. There then followed a series of increasingly strong Anglo-French ultimatums, and in September Venizélos proceeded to Salonika to set up a rival government. Although the Allies' terms were met on paper by King Constantine's successive governments, they grew steadily stiffer (especially after December 1, 1916, when Allied troops landing at Piraeus were fired upon by the King's troops as they approached Athens); and they culminated on June 11, 1917, in a demand that Constantine should leave the country, which he did the next day. His second son, Alexander, succeeded him. On June 26 Venizélos returned to Athens from Salonika as prime minister, and on June 29 Greece declared war on the Central Powers. Greek troops were not in action on the Macedonian front until May 1918, but they played a distinguished part in the final offensive in September and contributed an army corps to the French expedition into the Ukraine in December. Venizélos' subsequent triumph at the Conference of Paris was the climax of his achievement.

Succession of Alexander

As early as January 1915 the entente had tried to tempt Greece into the war with the promise of territorial gains

in Anatolia. Venizélos exacted fulfillment of this promise, and the Allied Supreme Council in Paris authorized the landing of Greek troops at Smyrna in May 1919. This was the beginning of disaster both for Venizélos and for Greece. He himself stayed away from Athens too long for his own good, achieving diplomatic triumphs: securing, for instance, the treaties of Neuilly and Sèvres, terminating the war with Bulgaria and with Turkey, respectively; reaching the agreement with Italy (July 1919) on the terms of cession of the Dodecanese to Greece; achieving an extension of the Greek area of occupation in Turkey (July 1919); and, not least, gaining the admiration of all the Allied representatives. But the greater part of this edifice collapsed within two years. In September 1920, on returning to Athens, Venizélos dissolved the chamber, in October King Alexander died, in November Venizélos was defeated at the polls, on December 20 Constantine was restored to his throne, and in the following year the occupation of Smyrna developed into a catastrophic war with Turkey led by Mustafa Kemal (later called Kemal Atatürk). The Greeks, ill-advised or imperfectly restrained by western European politicians, launched in January 1921 a general offensive in Anatolia, which was defeated and then, in July, obstinately renewed. By September they were in full retreat.

Restoration of Constantine

In October the prime minister, Gounaris, appealed to the British to intervene, while at the same time the French entered into an accord with Mustafa Kemal. The British government, being itself disunited on policy in the Middle East, was unable to rally its allies to a common policy. In August 1922 the Turks launched a final offensive that drove the Greeks out of Anatolia in September. The remnants of the Greek army rallied on the island of Chios under Gen. Nikolaos Plastiras, who put himself at the head of an antiroyalist revolution. Constantine left the throne to the crown prince George on September 27, and six of his principal ministers and generals (including Gounaris) were court-martialed and shot under Plastiras' orders.

The war with Turkey was ended by the Treaty of Lausanne on July 24, 1923, which superseded the unratified Treaty of Sèvres. An agreement for the compulsory exchange of Muslim and Orthodox populations between Greece and Turkey under the supervision of the League of Nations, following the similar Greco-Bulgarian agreement of 1919 (which, however, was on a voluntary basis), left Greece with a more homogeneous population but also with nearly 1,500,000 refugees to assimilate.

The population exchange especially affected Macedonia, where the largest proportion of immigrants settled. Macedonia, which had been 43 percent Greek, became 89 percent Greek. The refugees from Asia Minor had to endure a sharp drop in their standard of living; their arrival added to the existing overpopulation problem in often hostile surroundings. Land redistribution became absolutely necessary to provide the immigrants with a livelihood and, within a short time, the government reallocated nearly 38 percent of the arable land, taken from that vacated by Bulgarians and Turks and from church, state, and private holdings.

The country's troubles were aggravated by an incident on the Greco-Albanian frontier in August 1923, which led to an Italian bombardment of Corfu and to the imposition of a heavy indemnity by the League of Nations, and by an abortive military rising in Macedonia in October, the promotion of which was attributed to the retired chief of staff, General Metaxas, and other royalist officers.

General Plastiras' revolutionary committee accordingly asked the young king, George II, to leave the country while the future of the monarchy was decided. Anxious to restore constitutional government, the committee held elections on December 16, which were won by the Liberals. The committee then resigned, in January 1924, and after a brief tenure of office by Venizélos, a republic was proclaimed on March 25 and confirmed by a plebiscite on April 13. Adm. Pavlos Koundouriotis, scion of a family famous in the War of Independence and a hero of the First Balkan War, became president.

The republic, 1924–35. The new republic made an uneasy start, marred by friction with all its neighbours to the north and also with the major powers. Little more than a year after the republic's proclamation, and before it had a constitution, power was seized on June 26, 1925, by Gen. Theodoros Pangalos, one of the few officers who had been successful in the recent war with Turkey; on January 3, 1926, he assumed dictatorial powers. Koundouriotis resigned and was replaced by the dictator after a travestied election. Pangalos was an eccentric, though not a brutal, ruler but his methods were dangerous and had almost precipitated a war with Bulgaria in October 1925. He held power only until August 22, 1926, when he was deposed by another military coup d'état, led by Gen. Geórgios Kondílis.

Kondílis hastened to restore normal government. He dissolved the Republican Guard with some bloodshed in September, restored Koundouriotis to the presidency, organized elections in November, and, after a coalition government had taken office under Zaimis, retired from the scene. With some reconstruction, the same government held office until May 1928, and in the meantime a new constitution was promulgated on June 3, 1927. The government's fall was precipitated by the return of Venizélos, after several years' absence abroad, in March 1928. He became prime minister on July 3, dissolved the chamber a week later, and won a large majority at the general election in August.

Venizélos held office for nearly four years, during which time he was largely successful in restoring normal relations with all Greece's Balkan neighbours. The republic (of which Zaimis had been elected president in December 1929) had begun to appear stable and accepted. But beneath the surface there were disturbing currents.

Financial crisis of 1932

A financial crisis in 1932 weakened Venizélos' position. In May his government fell over two measures, which showed that the constitutional quarrel was not dead. One was to make representation in the electoral system proportional; the other was to restrict the freedom of the press, which the royalists, under the leader of the Popular Party, Panayiotis Tsaldaris, were allegedly abusing. In September, after two short-lived governments (the second again under Venizélos), there was a dissolution and a general election with proportional representation.

The result was to elect many small groups (including the Communists) with no clear majority for any party. Tsaldaris formed a precarious coalition that fell in January 1933. Another general election in March 1933 (by majority voting) enabled Tsaldaris to form a stable government, and it was this government that crowned Venizélos' work by signing a Balkan defense agreement with Romania, Turkey, and Yugoslavia on February 9, 1934. But by the following year a state of near deadlock had been reached between the two principal parties, Tsaldaris' Populists, who controlled the Chamber, and Venizélos' Liberals, who controlled the Senate. A battle was fought before the two agreed on the reelection of Zaimis as president in April 1934. The real issue was becoming increasingly plain: the Populists' intention to restore the monarchy.

Monarchy restored. In March 1935 an abortive coup d'état was staged with the object of putting Venizélos in power and frustrating the royalists. Its failure was the end of Venizélos' career and made the restoration of King George II certain. Kondílis, who put down the rising, was made deputy premier; after a general election in June, he took part in a coalition government of the right wing under Tsaldaris. After internal disputes about the method in which the monarchy was to be restored, Kondílis became first prime minister and then regent and organized a plebiscite on November 3, which approved the restoration of the monarchy. King George was back in Athens by the end of the month, and the constitution of 1911 was restored in place of that of 1927.

Restoration of George II

But the settlement was still uneasy. At a general election in January 1936 power was divided almost evenly between a Liberal and Republican group led by Themistoklis Sophoulis in succession to Venizélos and a right-wing coalition of the parties of Tsaldaris and Kondílis joined by a tiny group led by Metaxas, the former chief of staff. Between these two coalitions the balance was held by 15 Communists. A succession of deaths left Metaxas the only

possible prime minister, and on August 4, 1936, he made himself dictator, with the King's consent, in order to forestall a Communist coup d'état under cover of a general strike fixed for the following day. Metaxas' blow to the constitution and, especially, the King's acquiescence in it reopened bitter feelings between royalists and republicans, but probably saved Greece from anarchy.

World War II. "The Fourth of August," as Metaxas' regime was called, produced administrative efficiency, a sound currency, and adequate defenses, at the expense of the parliamentary system. It was hostile to the Communists but otherwise not unpopular. Metaxas had been expected by many to side with Germany and Italy in the world struggle already developing. In fact, he sought first to strengthen the Balkan entente inherited from Venizélos, which was extended by a nonaggression treaty with Bulgaria in July 1938. He accepted a British guarantee of Greece in April 1939, after Italy's annexation of Albania. On the morning of October 28, 1940, the Italian ambassador called on General Metaxas with his government's demand that Italian troops be allowed to occupy strategic points in Greece and the threat that any resistance would be crushed by force of arms. Metaxas is said to have answered curtly, "Oxi!" ("No"). His dismissal of Mussolini's representative and the Greek rout of the invading Italian army are celebrated annually on that date as Oxi Day.

German invasion of Greece

The Greek Army, under Alexandros Papagos, had occupied about a quarter of Albania before the Germans also attacked Greece (via Bulgaria and Yugoslavia in April 1941). Greece was then rapidly overrun; and a small British force, which had arrived in March 1941 in fulfillment of the guarantee of 1939, was driven off the mainland by the end of April and off Crete a month later.

By this time Metaxas had died (January 29), his successor had committed suicide, and King George II and his new government (under Emmanuel Tsouderos) had gone into exile. The occupation of Greece lasted to October 1944. It was marked by appalling suffering and great heroism. The Germans found successive dupes to be prime minister, but there were few others like them among the people. Resistance began with passive noncooperation, developed into sabotage and the formation of guerrilla bands in 1941–42, and culminated on November 25, 1942, with the destruction of the Gorgopotamos railway viaduct by a force of Greek guerrillas and British parachutists. This was the only occasion on which the two principal anti-German guerrilla forces, the Communist-controlled EAM-ELAS (National Liberation Front and National Popular Liberation army) and Col. Napoleon Zervas' EDES (Greek Democratic National army), cooperated in action. At other times they frequently fought each other, though not without undertaking independent actions against the common enemy, while the harbour they provided enabled the Allied liaison troops with them to conduct many activities that impeded the war effort of the Axis.

In a series of conflicts lasting from September 1943 (immediately after the surrender of Italy to the Allies) to the summer of 1944, EAM-ELAS eliminated all its political and guerrilla rivals except Zervas and set up a provisional government in the Greek mountains that by implication disowned both the King and his government in exile. A mutiny in sympathy occurred among the Greek troops in Egypt in April 1944 (when Tsouderos resigned); but after this had been suppressed and a precarious peace had been restored among the guerrillas, the two rival governments were brought together in a coalition under Georgios Papandreou, a former Venizélist Liberal. Accompanied by a small British force, his government returned to Athens as the Germans withdrew from Greece in October, but it disintegrated a few weeks later when the Communist members of the coalition refused to disband their guerrilla force. A bitter civil war broke out in Athens on December 3, which the British military forces suppressed with great difficulty, after EAM-ELAS had overrun virtually all Greece except Athens and Salonika.

Communist rebellion. The Communists accepted defeat and the disbandment of their forces at the Conference of Varkiza in February 1945. A period of political reconstruction followed, under the regency of the archbishop of Athens, Damaskinos, the King having agreed not to return to Greece pending a plebiscite on the monarchy. Many governments followed before an election was held in March 1946. The Communists abstained, and a royalist majority was returned. A plebiscite followed in September and restored George II. He died on April 1, 1947, and was succeeded by his brother Paul.

During 1946 a full-scale guerrilla war was reopened by the Communists. The commitment of defending Greece became too much for Great Britain, and it was taken on by the U.S. government with the announcement of the Truman Doctrine on March 12, 1947. Massive help from the U.S. came just in time, for by the end of 1947 large areas were in Communist hands again, and on December 24 the Communist leader Markos Vafiades proclaimed a provisional government in the northern mountains.

British and U.S. aid

Reconstruction and recovery. The rebellion did not end until 1949, after the defection of Yugoslavia from the Soviet bloc had closed a stretch of Greece's northern frontier to the rebels. On October 16, 1949, the Greek Communist broadcasting station announced the end of open hostilities. Reconstruction was an enormous job, one-tenth of the Greek population having died of war or starvation during the 1940s. The task was put in hand largely with U.S. funds and under U.S. guidance, but the continual change of governments inhibited progress. The election of March 1950 resulted in an unstable coalition of the Liberals and the left centre under Plastiras, and after frequent reshuffling of cabinets, another election was held in September 1951.

A few weeks earlier a new phenomenon had entered the political field. Papagos, who had led the Greek army against the Italians in 1940 and against the Communist rebels in 1949, resigned as commander in chief and formed a new political group, the Greek Rally (Ellinikos Synagermos). It won the largest number of votes, but not enough to enable Papagos to govern alone, and he refused to enter a coalition. The result was another year of unstable government until November 16, 1952, when another election was held, this time under the system of weighted proportional representation. The Rally obtained 239 seats out of 300, so that at last Greece was able to dispense with a coalition government. Papagos became prime minister and set about accelerating reconstruction. By the end of 1955, though still in need of U.S. assistance and in spite of a disastrous earthquake in 1953, Greece was well advanced in its recovery.

The Greeks could now assume a more considerable role in world affairs. The Dodecanese, ceded to Greece by Italy under the peace treaty of February 1947, were formally annexed in 1948. Greece entered the Council of Europe in 1949 and the North Atlantic Treaty Organization (NATO) in 1951. A Greek contingent took part in the Korean War (1950–53). On August 9, 1954, at Bled, Yugoslavia, Greece signed a treaty of alliance with its ancient enemy Turkey and with Communist Yugoslavia.

Following the death of Papagos on October 4, 1955, Konstantinos Karamanlis, who as minister of public works had won wide popularity, formed a new government and ensured some continuance of political stability. His position was confirmed at the election held on February 19, 1956, when women voted for the first time. The National Radical Union (Ethniki Rizospastiki Enosis [ERE]), a new party, won 165 seats in the 300-member Chamber of Deputies, the opposition being divided among seven different groups, including the crypto-Communists.

Government of Konstantinos Karamanlis

In March 1958 a political crisis arose through the resignation of two ministers and 13 deputies from the ERE because they disapproved of the government's intention to introduce a new electoral law that would favour the larger parties. Karamanlis thereupon resigned. In the subsequent general election, held on May 11, 1958, he was again returned, with 172 seats, but the Union of the Democratic Left (EDA), which had received the support of the outlawed Communist party, obtained 79 seats. It was clear that Communist competition was by no means a thing of the past.

When Karamanlis took over, a serious point of friction remained—the sovereignty of the British colony of Cyprus.

The campaign of Cypriot Greeks for union with Greece had been adopted by Papagos' government in 1954. The Greek claim soon spoiled harmony with Turkey, which thought that in the event of a British withdrawal the island should return to Turkey. Finally, by the Zürich agreement of February 1959, Greece and Turkey in effect agreed to Cypriot independence.

Throughout 1962 Georgios Papandreou's Centre Union campaigned against the government, disputing the validity of the election that returned Karamanlis in 1961. Papandreou also attacked King Paul, imputing interference by the crown in the government. Karamanlis resigned on June 11, 1963, and left the country. A caretaker government supervised the election of November 3, in which the Centre Union obtained 140 seats. Papandreou formed a new government but, failing to gain a vote of confidence, resigned. An election on February 16, 1964, however, gave him a clear majority (173 seats). Paul I died on March 6, 1964, and his son Constantine succeeded.

On July 15, 1965, Constantine dismissed Papandreou from office, supposedly for aiding penetration of the army by the left-wing organization Aspida, allegedly headed by his civilian son, Andreas Papandreou. After withholding confidence from two successive governments, the Chamber on September 25, 1965, finally approved one formed by Stephanos Stephanopoulos, the candidate of 45 dissidents from the Centre Union; with the votes of 99 deputies of the ERE and eight Progressives, it held a majority of four.

As regards Aspida, 28 officers were indicted in October 1966; but, while their trial was in progress, Panayotis Kanellopoulos, leader of the ERE, withdrew support from the government, and Stephanopoulos resigned on December 21. The King promptly asked Ioannis Paraskevopoulos, a banker, to form a caretaker government with a view to holding an election in May 1967. Paraskevopoulos, however, had to resign office on March 30, and a government under Kanellopoulos was formed on April 3.

Military government, 1967–74. Kanellopoulos dissolved Parliament on April 14, but on April 21, before elections could be held, a military coup d'état overthrew his government and transferred power to a group of army officers, acting in the King's name but without the King's knowledge. A new Cabinet was installed with Konstantinos Kollias, chief prosecutor of the Supreme Court, as prime minister on the King's insistence, but with three military men, Lieut.-Gen. Gregorios Spandidakis, Brig. Stylianos Patakos, and Col. Georgios Papadopoulos, in key posts.

The government arrested large numbers of politicians, including Georgios and Andreas Papandreou, and eventually Kanellopoulos, as well as many alleged Communists. Though establishing strict censorship and governing unconstitutionally, it announced in May 1967 that a new constitution would be drafted and submitted for a referendum. Meanwhile, Andreas Papandreou was tried and sentenced in August to nine years' imprisonment for attempting to overthrow the regime. That autumn there were several purges of the armed forces, the civil service, and academics, and in December trial by jury was abolished.

On December 13 the King broadcast from Lárisa an appeal for the armed forces and the people to overthrow the junta. The appeal failed, and the King and his family fled to Rome, where he stated that he had not opposed the coup in April because to have done so would have caused bloodshed and that he had hoped by remaining in Greece to persuade those in power to restore democracy; his only condition for returning was that there should be a timetable for the reestablishment of democracy. The government immediately appointed Gen. Georgios Zoitakis regent, and, relinquishing military rank, Papadopoulos and Patakos took over as prime minister and deputy prime minister, respectively. An amnesty was proclaimed on December 23 for all political prisoners, including Andreas Papandreou, who then went abroad. Georgios, his father, had been released shortly before, though later he was put under house arrest; he died in November 1968.

Constitution of 1968

The new constitution abolished the remaining prerogatives of the king and provided for a unicameral parliament of 150 deputies (300 under the previous constitution) elected by secret ballot, with obligatory voting. It was approved on September 29, 1968, by a national referendum, after a vigorous campaign on its behalf in the censored press. Political activity was prohibited, and the junta ruled by decree. Under the new constitution, King Constantine could return to his throne only after the first parliamentary election, unless the government decided to recall him earlier. The regime stated, however, that a general election would be held only after the "Revolution of April 21" had proved successful in reforming the "Greek mentality."

In January 1968 the Consultative Assembly of the Council of Europe resolved that Greece should be expelled from membership if by 1969 there had been no return to parliamentary democracy. In December 1969 Greece withdrew from the Council of Europe.

Opposition to the regime was organized from abroad, one of its creators being Andreas Papandreou. In July and August 1971 the regime deprived many prominent expatriates of their Greek nationality for "engaging in antinational activities abroad."

On March 21, 1972, Zoitakis was dismissed and Papadopoulos took over as regent. On June 1, 1973, the junta decreed that the monarchy was abolished and replaced by a republic. This was confirmed by popular referendum on July 29, but the validity of this referendum was challenged. In July, Papadopoulos was appointed president, martial law ended, and plans were made for the coming elections. After the fall of the junta at the end of 1973, effective power still lay with a military clique. Another referendum was held, on December 8, 1974, and 69.2 percent of those voting chose "uncrowned democracy."

The regime's underlying problems continued to plague the government. "A way out," in the form of a glorious foreign policy accomplishment, seemed to offer itself to desperate men, and so the Athens government permitted, if it did not plan and direct, the overthrow in Cyprus in July 1974 of the established government of Archbishop Makarios, with a view to bringing Cyprus into union (*enosis*) with Greece. The Turkish government, eagerly invoking its role as guardian of the Turkish minority (18 percent of the population) on Cyprus, responded on July 20 with an invasion force that occupied about 40 percent of the island and turned almost 200,000 Greek Cypriots into refugees. War threatened between Turkey and Greece.

Restoration of parliamentary government. At this point the Cabinet disintegrated, and Konstantinos Karamanlis, a former prime minister, was called from exile to occupy the post again and form a civilian "Government of National Salvation." On August 1, 1974, Karamanlis announced that the constitution of 1952 would be reintroduced as a provisional measure until, in his words, "the country acquires a charter fully approved by the people." The powers of head of state were to be exercised by the president of the republic. The word king in the constitution was replaced by the word president. Under the restored constitution of 1952 the government had full control over the armed forces, and the judiciary was free.

There was a general loosening of civil restrictions, and, in September 1974, rival parties began to emerge. Many Greeks who had expatriated for political reasons were able to return to Greece. Parliamentary elections were held on November 17, 1974, the first since 1964. The New Democracy Party of Karamanlis won 54 percent of the votes cast and a very large majority of the seats in Parliament; the next largest number of seats went to the Centre Union–New Forces Party. On December 9, 1974, the Greek Parliament met for the first time in seven years. A November 1975 election again showed strong support for Karamanlis' New Democracy Party.

The draft of a new constitution providing for a parliamentary democracy was published on December 23, 1974, after endorsement by the full Cabinet. The constitution was promulgated in June 1975, and Konstantinos Tsatsos was elected president. In the following years, the New Democracy Party began to decline. Karamanlis became president and Georgios Rallis took over as prime minister. In the 1981 elections, however, the Pan-Hellenic Socialist Movement (Pasok) gained an absolute majority. It thus placed its leader, Andreas Papandreou, as prime minister in Greece's first Socialist government.

Greece became the 10th full member of the European Economic Community (EEC) in January 1981. Its membership in the EEC and in NATO was opposed by several Greek political factions, as well as many in Pasok. Although Papandreou was critical of the U.S. military bases in Greece, in 1983 he signed an agreement that allowed the bases to remain in operation. Throughout the 1980s the Cyprus controversy, reinforced by disputes over control of the Aegean Sea, continued to embitter relations between Greece and Turkey and between Greece and the United States.

In early 1985 Papandreou unexpectedly withdrew support for Karamanlis' reelection, forcing the President to resign. Parliament replaced him with Papandreou's nominee, Christos Sartzetakis, a Supreme Court judge. A series of constitutional amendments passed in 1986 severely limited the president's powers.

While Pasok won a majority of seats in the 1985 parliamentary elections, its popularity declined in the late 1980s, in part from the government's economic austerity program and from a series of scandals involving government corruption. In June 1989 no party received a majority in elections to Parliament, and the conservative New Democracy Party and the Communist-controlled Left Coalition agreed to a short-term alliance in order to initiate criminal proceedings against a number of Pasok ministers, including Papandreou, for corruption and fraud. Because elections held in November were also indecisive, an all-party coalition was established to run the government until new elections, scheduled for April 1990. The New Democracy Party won exactly half the parliamentary seats in these elections, and its leader, Constantine Mitsotakis, became prime minister. Parliament returned the 83-year-old Karamanlis to the presidency.

For later developments in the history of Greece, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 912, 921, 961, 963, and 972, and the *Index.*

(B.S.-E./C.M.Wo./D.V.D./J.M.H./
N.G.L.H./W.Mr./S.P.A./G.An./Ed.)

BIBLIOGRAPHY. A good source for readings on Greece is MARY JO CLOGG and RICHARD CLOGG (comps.), *Greece* (1980), a bibliography with more than 800 entries on some 30 subjects, with the majority of cited sources in English. Discussions of Greek affairs and character are found in GEORGE ANASTAPLO, *Human Being and Citizen* (1975), and *Artist and Thinker* (1983); see also his *Constitutionalist* (1971), which includes citations to 10 of the author's articles on contemporary Greek affairs republished in the *Congressional Record.*

Physical and human geography: The two most extensive works on the geography of Greece are ALFRED PHILIPPSON, *Die griechischen Landschaften,* 4 vol. (1950–59); and *Synchronē geōgraphiaatlas tes Hellados,* ed. by PERIKLES RODAKES and KOSTAS TRIANTAPHYLLOU, 2 vol. (1960–69). In English, GREAT BRITAIN. NAVAL INTELLIGENCE DIVISION, *Greece,* 3 vol. (1944–45), is still the best summary of conditions to that time; later events and trends may be studied in JOHN CAMPBELL and PHILIP SHERRARD, *Modern Greece* (1968). For publications providing the statistical basis for analysis of Greece's population, society, and economy, see works published by the BANK OF GREECE and by the NATIONAL STATISTICAL SERVICE OF GREECE (chief of which is the annual *Statistical Yearbook of Greece*), as well as the publications of various European organizations, especially the ORGANISATION FOR ECONOMIC CO-OPERATION AND DEVELOPMENT (OECD), especially its series *Economic Surveys.* Comprehensive guidebooks include STUART ROSSITER, *Greece,* 3rd ed. (1977, reprinted 1980), in the series *Blue Guides;* ROBERT BOULANGER, *Grèce,* rev. and updated ed., ed. by JEAN-JACQUES MAFFRE, PHILIPPE LANNOIS, and PATRICE MILLERON (1981); and OTTO GARTNER and GERALD SAWADE, *Baedeker's Greece* (1981; originally published in German, 1979).

The land: In addition to the general geographical works cited above, reference should be made to *Geological and Physicogeographical Bibliography of Greece* (1961), by D. HARALAMBOUS; specific publications dealing with individual aspects of the landscape include E.G. MARIOLOPOULOS, *An Outline of the Climate of Greece,* trans. from the Greek (1961).

The people: The monograph *Population of Greece* (1974), prepared by DIMITRIOS TRICHOPOULOS and GEORGE PAPAEVANGELOU for the World Population Year 1974, gives an authoritative picture of Greece's population both in modern times and in historical perspective; other useful studies include BERNARD KAYSER, *Géographie humaine de la Grèce* (1964), and D. PENTZOPOULOS, *The Balkan Exchange of Minorities and Its Impact upon Greece* (1962). Publications of the censuses of 1961, 1971, and 1981, by the NATIONAL STATISTICAL SERVICE OF GREECE, provide both basic data and analysis.

The national economy: For characteristics of Greece's people and economy, see BERNARD KAYSER and KENNETH THOMPSON, *Economic and Social Atlas of Greece* (1964). Useful surveys include *Problems of Greek Regional Development* (1962), by BENJAMIN WARD; and *Long-Term Prospects for the Greek Economy: A Forecast of Development in the Next 15 Years* (1968), issued by the ROYAL HELLENIC RESEARCH FOUNDATION.

Government and social conditions: Changing Greek social conditions and the relatively unchanging Greek countryside, people, and customs are described in PATRICK LEIGH FERMOR, *Mani: Travels in the Southern Peloponnese* (1958); TIMOTHY WARE, *The Orthodox Church* (1964, reprinted 1983); ERNESTINE FRIEDL, *Vasilika: A Village in Modern Greece* (1962); JOHN K. CAMPBELL, *Honour, Family and Patronage: A Study of Institutions and Moral Values in a Greek Mountain Community* (1964); BERNARD KAYSER, PIERRE-YVES PECHOUX, and MICHEL SIVIGNON, *Exode rural et attraction urbaine en Grèce* (1971); and PETER LEVI, *The Hill of Kronos* (1981). WILLIAM H. MCNEILL, *The Metamorphosis of Greece Since World War II* (1978); and ROBERT MCDONALD, *Pillar and Tinderbox: The Greek Press and the Dictatorship* (1983), focus on political developments.

Culture: See JOHN BOARDMAN, *Greek Art,* rev. ed. (1973); CONSTANTINE A. TRYPANIS, *Medieval and Modern Greek Poetry* (1951, reprinted 1964), an anthology of poetry in Greek, with a good introduction in English, and his *Penguin Book of Greek Verse* (1971, reprinted 1979). See also KIMON FRIAR (comp.), *Modern Greek Poetry* (1973).

History: (*From 1200 to 500 BC*): N.G.L. HAMMOND, *A History of Greece to 322 B.C.,* 2nd ed. (1967, reprinted 1981), gives a full account of political history of the period. MICHAEL GRANT (ed.), *The Birth of Western Civilization: Greece and Rome* (1964), is a collection of essays on Greek history, art, and thought. A wealth of information may be found in *The Cambridge Ancient History,* a multivolume work continuously updated and revised. PAUL MACKENDRICK, *The Greek Stones Speak,* 2nd ed. (1981), is a work on Greek archaeology. See also translations of the works by the ancient Greeks themselves: Herodotus, Thucydides, Diodorus, Arrian, Xenophon, and Plutarch. (*From 500 to 323 BC*): In addition to the ancient writers mentioned above, modern works include A.R. BURN, *Persia and the Greeks: The Deference of the West, c. 546–478 B.C.,* 2nd ed. (1984); DONALD KAGAN, *The Outbreak of the Peloponnesian War* (1969); and J.B. BURY and RUSSELL MEIGGS, *A History of Greece to the Death of Alexander the Great,* 4th ed. (1975). PIERRE LÉVÊQUE, *The Greek Adventure* (1968; originally published in French, 1964); and FRANÇOIS CHAMOUX, *La Civilisation grecque à l'époque archaïque et classique,* rev. ed. (1983), are complementary works dealing with Greek civilization in the largest sense. (*The Hellenistic Age*): For a general approach, consult SIR WILLIAM TARN, *Hellenistic Civilisation,* 3rd ed., rev. by the author and G.T. GRIFFITH (1952, reissued 1975). Useful information is found in ÉDOUARD WILL, *Histoire politique du monde hellénistique: 323–30 av. J.-C.,* 2nd rev. ed., 2 vol. (1979–82); and A.R. BURN, *Alexander the Great and the Hellenistic World,* new ed. (1962). (*The Roman and Byzantine periods*): PETER BROWN, *The World of Late Antiquity, AD 150–750* (1971), is an excellent cultural history; SPEROS VRYONIS, *Byzantium and Europe* (1967), is a representative volume in the series *Library of European Civilization;* DIMITRI OBOLENSKI, *The Byzantine Commonwealth: Eastern Europe, 500–1453* (1971, reissued 1982), discusses the Balkan region as a whole; see also *The Cambridge Medieval History,* vol. 4: *The Byzantine Empire,* parts 1 and 2, new ed. (1966–67). (*Ottoman period*): L.S. STAVRIANOS, *The Balkans Since 1453* (1958), provides a general discussion. (*From 1815 to 1914*): L.S. STAVRIANOS, *The Balkans, 1815–1914* (1963, reprinted 1974), is a short review; NICOLAS G. SVORONOS, *Histoire de la Grèce moderne,* 3rd ed. (1972), is a short but excellent history; see also EDWARD S. FORSTER, *A Short History of Modern Greece, 1821–1956,* 3rd ed., rev. by DOUGLAS DAKIN (1958, reprinted 1977). (*After 1914*): See C.M. WOODHOUSE, *The Story of Modern Greece* (1968); W.A. HEURTLEY *et al., A Short History of Greece from Early Times to 1964* (1965); ALEXANDER ELIOT, *The Horizon Concise History of Greece* (1972); and RICHARD CLOGG, *A Short History of Modern Greece* (1979).

Ancient Greek and Roman Civilizations

Classical Greco-Roman civilization is an overall term that refers to Aegean and ancient Greek civilizations, ancient Italic peoples, and ancient Roman civilization, all of which flourished on the continents of Europe and Asia from the Neolithic Period (New Stone Age) to the decline of the Western Roman Empire in the 5th century AD and are treated here.

The article is divided into the following sections:

Aegean civilizations 205
- Early Aegean civilizations 207
  - Early cultures
  - The Bronze Age
- The decline of the early Aegean civilizations 212
  - The eruption of Thera (*c.* 1500) and the conquest of Crete (*c.* 1450)
  - The end of the Bronze Age in the Aegean (*c.* 1200–1000)
  - The people of the Bronze Age Aegean

Ancient Greek civilization 216
- The Early Archaic and Archaic periods 217
  - The Early Archaic period
  - The Greek colonizing movement, *c.* 750–*c.* 500 BC
  - Archaic Greek culture
- The Greek city-state 225
  - The growth of the *polis*
  - Sparta
  - Athens
  - Other important *poleis*
  - Relations among the Greek *poleis* to the end of the 6th century BC
- The 5th century BC 240
  - The period of the Greco-Persian Wars
  - The Athenian empire
  - The Great Peloponnesian War (431–404 BC)
  - The cultural life of Athens during the war
- The Greek world from 404 to 323 BC 253
  - The Spartan hegemony and decline
  - The rise of Macedonia
  - The western Greeks
  - Greek culture and society in the 4th century BC

The Hellenistic Age 265
- The conquests of Alexander the Great 265
  - Relations between the Greeks and Macedonians
  - The conquest of the Persian Empire (334–330)
  - The conquests in Bactria and the Indus Valley (330–323)
  - The empire and achievements of Alexander
- The Hellenistic Age (323–30 BC): Greece in the Roman period 268
  - The establishment of the Hellenistic kingdoms (323–276 BC)
  - The Hellenistic monarchies in the 3rd century BC
  - Relations among the Hellenistic states, 275 to 30 BC
  - Hellenistic political, social, economic, and cultural institutions
  - Macedonia and Greece under Roman rule (to *c.* AD 395)

Ancient Italic peoples 282
- The Etruscans 282
  - Origins
  - Language and writing
  - Characteristics
  - Organization
  - Aspects of Etruscan civilization
  - Etruscan expansion and foreign relations
  - The Roman conquest of Etruria
- Other Italic peoples 288
  - Local populations in areas colonized by Greece
  - The Apulians
  - The Latins
  - The eastern Italics
  - The populations of the Picenum
  - The Veneti
  - The Ligurians
  - Populations of central northern Italy and of the Alps

Ancient Rome 290
- Rome from its origins to 264 BC 290
  - Early Rome to the 6th century BC
  - Early centuries of the Roman Republic (6th century BC–264 BC)
  - The expansion of Rome in Italy
- The middle republic (264–133 BC) 296
  - The first two Punic Wars
  - The establishment of Roman hegemony in the Hellenistic world
  - Roman government, economy, and culture
- The late republic (133–31 BC) 306
  - The aftermath of the victories
  - The reform movement of the Gracchi (133–121 BC)
  - The republic (*c.* 121–91 BC)
  - Wars and dictatorship (*c.* 91–80 BC)
  - The Roman state in the two decades after Sulla (79–60 BC)
  - The final collapse of the Roman Republic (59–44 BC)
  - The Triumvirate and Octavian's achievement of sole power
- The early Roman Empire (31 BC–AD 193) 313
  - The consolidation of the empire under the Julio-Claudians
  - Growth of the empire under the Flavians and Antonines
  - The empire in the 2nd century
- The later Roman Empire 326
  - The dynasty of the Severi (AD 193–235)
  - Religious and cultural life in the 3rd century
  - Military anarchy and the disintegration of the empire (235–270)
  - The recovery of the empire and the establishment of the dominate (270–337)
  - The Roman Empire under the 4th-century successors of Constantine
  - The eclipse of the Roman Empire in the West (*c.* 395–500) and the German migrations

Bibliography 338

Aegean civilizations

The adjective that is traditionally used to name those Bronze Age civilizations that arose and flourished in the area of the Aegean Sea between *c.* 3000 and 1000 BC is Aegean.

The area consists of Crete, the Cyclades and some other islands, and the Greek mainland, including the Peloponnese, central Greece, and Thessaly. The first high civilization on European soil, with stately palaces, fine craftsmanship, and writing, developed in the island of Crete. Later, the peoples of the mainland adapted the Cretan civilization to form their own, much as the Romans adapted the civilization of later Greece. The Bronze Age civilization of Crete has been called Minoan, after the legendary King Minos of Knossos, which was the chief city of the island throughout early times. The Bronze Age of the Cyclades is known as Cycladic, that of the mainland as Helladic, from Hellas, the Greek name for Greece. Early, middle, and late stages have been defined in each of these, with further subdivisions according to recognizable changes in the style of pottery and other products that are associated with each separate culture. The civilization that arose on the mainland under Cretan influence in the 16th century BC is called Mycenaean after Mycenae, which appears to have been its most important centre. The term Mycenaean is also sometimes used for the civilizations of the Aegean area as a whole from *c.* 1400 BC onward.

Dating of the Aegean Bronze Age. The dates that are suggested here are approximate and conventional. In a general way, they are based on correlations with Egypt, where, from the beginning of the Dynastic Period (*c.* 3100 BC onward), a historical chronology can be established with a leeway of a few centuries and can be fixed within



Principal sites associated with Aegean civilizations.

reasonably narrow limits after *c.* 2000 BC. Bronze Age pottery from the Aegean has been found in contexts in Egypt that are datable, and many Egyptian objects have been recovered in the island of Crete.

Two important landmarks are fragments of Cretan pottery from the town at Kahun in the Fayyūm, built for workers who were engaged in the construction of a pyramid for the pharaoh Sesostris II (1897–78), and a large quantity of Mycenaean pottery from the mainland found at Tell el-Amarna, site of Akhenaton's capital, and imported during his reign (1379–62). Radiocarbon dates appear consistent with those based on correlations with Egypt.

History of exploration. The poems of Homer, although they reflect Greek society in the 8th century BC, describe warriors using obsolete bronze weapons and objects such as helmets plated with tusks of wild boar that went out of use before the end of the Aegean Bronze Age. Massive Bronze Age defense walls survived at Mycenae and elsewhere on the mainland and were called Cyclopean because, according to Greek tradition, the Cyclopes had built them. Apart from these Cyclopean walls, virtually nothing was known about the Aegean Bronze Age before the middle of the 19th century, when, in 1876, a German archaeologist, Heinrich Schliemann, discovered unplundered royal shaft graves at Mycenae and thought that the men buried in them were the Greek heroes of Homer's siege of Troy. Actually, the graves spanned the 16th century BC and proved to be older than any likely date for a Trojan war. Schliemann's discoveries, however, led to intensive exploration of Bronze Age and earlier sites on the Greek mainland. Meanwhile, in the 1860s, Bronze Age settlements of the period of the Mycenae shaft graves had come to light under the debris of an immense eruption on the volcanic island of Thera (Thíra). Later in the 19th century, Christos Tsountas, a Greek archaeologist, dug cemeteries of earlier phases of the Bronze Age on other Cycladic islands and continued the work begun by Schliemann at Mycenae. At the end of the century, a British expedition excavated an important Bronze Age town at Phylakopi on Melos (Mílos). When Crete eventually became independent of Turkish rule in 1898, attention was turned to Bronze Age sites there. In 1900 Arthur (later Sir Arthur) Evans, an English archaeologist, began to uncover the palace at Knossos, the largest Bronze Age centre of the island, discovering clay tablets with the first positive evidence for Bronze Age writing in the Aegean. Greek, U.S., French, and Italian excavators added further knowledge of the Cretan Bronze Age during the years that followed, and U.S. and German expeditions opened new sites on the mainland. Inscribed clay tablets such as those found at Knossos in Crete at the turn of the century were recovered in Messenia; others have since come to light at Mycenae and elsewhere on the mainland. The belief that the language of these tablets might be Greek was established by an English archaeologist, Michael Ventris, in 1952, though acceptance of this is not yet universal. In 1962 a large palace, destroyed by fire *c.* 1450 at Zákros (Zakro) in eastern Crete, was found. In 1967 the clearing of a major settlement that was buried by the eruption of Akrotíri, on Thera, began.

The work of Schliemann

EARLY AEGEAN CIVILIZATIONS

Early cultures. *Paleolithic (Old Stone Age).* Chipped stone tools made by Paleolithic hunters have been found in many parts of mainland Greece, but none are yet recorded from Crete or the other islands. As elsewhere in Europe, the latest Lower Paleolithic industries evolved into Upper Paleolithic ones with diminutive stonework. During this Upper Paleolithic stage, obsidian, a volcanic glass, was brought from the island of Melos for making implements, indicating that boats capable of sea voyages were already in use in the Aegean area by then.

Neolithic (New Stone Age). If radiocarbon dates are to be trusted, agriculture was being practiced in some parts of the Aegean area as early as the 7th millennium. The first agriculturalists in the Aegean, like those of Anatolia and Palestine, may have been ignorant of the art of making fired clay vases—traces of agricultural settlements without pottery have been identified at several places in Thessaly and at Knossos in Crete. The island of Crete appears to have been uninhabited before this time, and the first agriculturalists must have reached it by sea from western Anatolia or from somewhere more distant, such as Cilicia. Other immigrants from the east may have brought agricultural techniques and ways of life to the mainland, where they mingled with the Upper Paleolithic hunting peoples whom they found there. The Aegean is one of the most favoured regions of the Mediterranean basin. Immigrants from the coastal areas of Anatolia, Syria, or Palestine would have found the climate and ecology similar to what they had known in their homelands. The olive and vine, sources of oil and wine, the staples of Mediterranean diet, grow in most parts of the Aegean area and may have been native there. Water, which is a problem in the present century, was probably more abundant in early times when forests were more extensive than they are now.

Ecology of the Aegean

Agricultural communities were eventually established in every part of Greece, making pottery by hand and grinding stones to shape edged tools, axes, adzes, and chisels. Wheat and barley were among crops grown. Hunting and fishing still played their part in obtaining food alongside raising of cattle, sheep, goats, and pigs. Arrowheads of chipped stone were used on the mainland and in the Cyclades, but none is recorded from Crete, where bone points may have served to tip arrows. Another long-range weapon was the sling, and clay sling pellets were made in Thessaly where suitable beach pebbles were not to hand. In Crete, clubs were armed with stone heads as in Egypt and elsewhere in the Near East in early times. Houses with rectangular rooms are attested at Knossos in Crete, at Saligos in the Cyclades, and in Thessaly. But some Aegean communities may have lived in circular huts of the kind found in Predynastic Egypt and in early Syria and Cyprus. Caves appear to have been used as cemeteries in Crete. Some cremations are assigned to the end of the Neolithic on the mainland. By that time, to judge from a house, larger than the rest, occupying a central position at Dhimíni (Dhimínion) in Thessaly, some communities may have been organized under chiefs. Several Thessalian settlements were surrounded by defense walls or ditches. Copper tools—simple, flat axes and knives—were in use before the end of the Neolithic both in Crete and on the mainland. Indeed, some of the pottery shapes characteristic of later phases of the Neolithic in Thessaly appear to be copied from those of metal vases. But this Thessalian late Neolithic may actually overlap in time with the beginning of the Bronze Age farther south.

The Bronze Age. *The Early Bronze Age (c. 3000–2200).* The transition from Neolithic to Bronze Age in the Aegean was marked by changes in pottery and other aspects of material culture. These may reflect the arrival in Crete and the Cyclades of new people from more civilized lands farther east bringing knowledge of metalworking with them. In Crete and the islands the changes that inaugurated the Bronze Age were more or less contemporary with the beginning of Dynastic times in Egypt. What is defined as the Bronze Age in the Peloponnese appears to have begun later under the influence of settlers from the islands. The Bronze Age in central Greece and Thessaly may have begun later still. Evolved types of metal tools appear to have been current considerably before the end of the Neolithic there.

Flourishing metal-using cultures were established by the middle of the 3rd millennium in Crete, the Cycladic islands, and the southern part of the mainland. Each of these three cultures had its own distinctive characteristics; but they shared much in common, and their peoples may have spoken the same or similar non-Greek languages. Many place-names throughout the Aegean, notably ones ending in *-nt-* and *-ss-*, such as Corinth and Knossos, seem to reflect a time when the same language (or a group of related languages) with Anatolian affinities was spoken there before the introduction of Greek. A large number of words came to be adopted into Greek from this earlier language group.

These Early Bronze Age peoples of the Aegean seem to have employed similar types of metal tools, including axes, adzes, and short daggers, but double axes may have been special to Crete. Tweezers were used for plucking facial hairs and rectangular stone palettes for grinding face paints with small pestles made of attractive veined stones or Spondylus shell.

Early metal tools

Lerna and other settlements on the mainland were eventually surrounded by massive walls with projecting towers, but whether similar defenses existed in the Cyclades and in Crete is uncertain. Houses with several rooms were being constructed in most parts of the Aegean by this time, and buildings at Knossos and at Vasilikí in Crete have been identified as the residences of local rulers. The so-called House of Tiles at Lerna, destroyed by fire toward the end of the period, appears to have been another of these "palaces." It was a large rectangular building with two stories taking its name from the baked clay tiles found in its ruins. These small, flat tiles are thought to have come from a sloping roof and may be the earliest roof tiles known. Similar tiles were recovered from a huge circular structure of the same period at neighbouring Tiryns, of which only a section has been excavated, as it lies deep below the level of the later Mycenaean palace there. It was evidently a public building of some kind, a meeting place, perhaps, or possibly even a temple. Some shrines of this period in Crete may have been circular, as were many of the early Cretan tombs.

But circular buildings at Orchomenus in central Greece are interpreted as granaries or huts. A remarkable stone vase of this period, from Melos, is in the shape of a granary complex with round silos, such as the granary of the later palace at Mallia (Mália) in Crete. The vase shows an entrance porch with a gabled roof, but it would appear that the roofs of houses in the Cyclades and in Crete were mostly flat at this time and later.

Cretans in the Early Bronze Age buried their dead in communal tombs. These belonged to clans or extended families and might remain in use for many generations. Traces of hundreds of burials have been noted in some of them. Caves and rock-shelters, as well as buildings of various kinds, were used as tombs. Circular tombs were characteristic of the Mesara (Mesarás) region of southern Crete. They were built above ground, and none has survived complete; but the massive walls that are left seem designed to support the weight of vaults, which may have been constructed of rough fieldstones held together with mud, as were the walls. Some of the largest tombs, however, with a diameter of 40 feet (12 metres) or more inside, may have been vaulted in mud brick. Annexes with cult rooms were built in front of the entrances of some tombs, and others had chambers for offerings around the sides. When a tomb became full, a new floor was laid above the earlier burials, or parts of the tomb's annex were brought into use as burial chambers. Sometimes the remains of earlier burials were removed to separate buildings or enclosures nearby. These circular tombs may have been introduced to Crete by people whose ancestors had lived in round huts. The inhabitants of the Neolithic settlement at Khirokitia in Cyprus, for instance, buried their dead under the floors of the round huts in which they lived. Communal tombs in some parts of Crete had rectangular compartments or rooms and flat roofs, such as those in contemporary houses. At Knossos, where the

Cretan burial practices

local rock was soft, artificial caves were dug as tombs. Everywhere in Crete the dead were normally trussed into a tight contracted position, knees to chin. Sometimes the bodies were then squeezed into large storage jars or small clay chests or coffins. There was evidently much ceremonial in connection with burial, and, apart from objects of personal use, such as seals, jewelry, and weapons, left with the dead, vases with offerings were regularly placed inside or outside the tombs.

In contrast to the Cretans, the people of the Cyclades during the earlier part of the Bronze Age buried their dead in small graves that held a single body, sometimes a pair, and never more than six or seven. The graves were often grouped in family cemeteries, which might be surrounded by a wall. The bodies were placed in them lying on one side in a loosely flexed position. Some Cycladic graves were little, stone-built chambers with an entrance, although the standard type consisted of a box (cist) made with large slabs set on edge and roofed with slabs.

Less is known about contemporary burials on the mainland, but the graves there normally contained several bodies, which suggests that they belonged to families but not to large units, such as the clans that existed in Crete. Various types of mainland graves of this period are known, including chambers cut in the rock and stone-built tombs, such as those in the Cyclades. Circular cairns, each covering several burials, on the island of Leucas in western Greece appear to go back to this time.

Early pottery

Pottery was still made by hand throughout the Aegean area. A useful type of vase first attested there at the beginning of the Early Bronze Age was a handled jug with a spout for pouring. Some of the earliest jugs from Crete have round bottoms and yellowish surfaces, as if they were copies of vessels made from gourds. Distinctive spouted bowls of oval shape that may have been drinking cups but are nicknamed sauce boats were quite typical of the Early Bronze Age on the mainland and usually have a fine reddish or dark overall wash. Pottery with a similar wash and with the surface often deliberately mottled is found in Crete and is known as Vasilikí ware, after a site with a little "palace" where large amounts of it were recovered. The pottery of the Cyclades during this period shares features in common with that of Crete and the mainland. The art of making stone vases flourished in the Cyclades from the beginning of the Bronze Age and may have spread from there to Crete, but it was hardly practiced on the mainland after the end of the Neolithic. The techniques used were simple and included boring with a hollow reed, which twirled an abrasive, either emery from Naxos or sand. The people of the Cyclades used their fine white marble not only for vases but for remarkable figurines, mostly female but including men, some playing double pipes or seated on chairs with harps. While the majority of these figures are only a few inches high, some females are larger and a few are nearly life-size. Some have traces of painted decoration. These marble figures were often placed in graves, and groups of them have been found in sanctuaries, though whether they represented gods and goddesses is uncertain. They were exported to the mainland and Crete and may have been imitated there. Vessels of gold and silver were current in the Aegean by now, and a few have survived, including gold sauce boats of mainland type and gold and silver bowls from the islands. Gold and silver jewelry of this period, mostly from Crete, includes bracelets, necklaces, earrings, headbands, and hair ornaments of various kinds. Some of the finest of this early jewelry was found in communal tombs at Mochlos on the northern coast of Crete. The inspiration for it no doubt came from the east, and much of that from Mochlos, notably hairpins with flower heads, is reminiscent of jewelry from the royal tombs at Ur in Mesopotamia.

Protective seals

Seals began to come into use in the Aegean about the middle of the 3rd millennium. In an age before the invention of locks and keys, seals were employed to stamp wet clay, which was used to secure doors or affix lids to storage jars or other containers. The design impressed by the seal might add the threat of magic to that of detection if the sealing was broken. Many seals resembling those current in Egypt and Syria have been recovered from early tombs in Crete. Some of these early Cretan seals were made of ivory, possibly obtained from Syria, where elephants still abounded. Others were made of bone or soft, easily cut stones, such as serpentine and steatite. They were of various shapes, some of animals or birds or their heads, others cylindrical, adapted from Syrian versions of early Mesopotamian cylinder seals. They were engraved with a variety of designs, including abstract patterns and pictures of animals, notably dangerous lions and scorpions or poisonous spiders (rogalidhas) of a species native to Crete. Seals appear to have been in use in the Cyclades and on the mainland during this period, but very few have been recovered there. A stone cylinder seal from Amorgos resembles early Syrian ones. Most Aegean seals of this period, however, even in Crete, may have been made of perishable wood. Clay-seal impressions, preserved by fires that destroyed buildings at Lerna, including the palatial House of Tiles, look as if they had been stamped by wooden seals. The designs on these Lerna impressions resemble those found on seals in Crete.

Pictures of boats with many oars or paddles were scratched on clay vases of this period in the Cyclades and on the mainland. The boats have a high prow often surmounted by a fish ensign, the stern being low with the keel apparently projecting beyond it. Similar vessels, though with a single mast for a square sail in addition to oars, are represented on early Cretan seals. Ships of this kind would have been capable of voyages to Syria and Egypt, whence skills and fashions were reaching Crete along with imports such as Egyptian stone vases and Syrian daggers.

End of the Early Bronze Age on the mainland (c. 2200–2000). The comparative unity of incipient civilization in the Aegean area was eventually shattered by new movements of people into the Cyclades and the southern part of the mainland. Toward the end of the 3rd millennium, many of the settlements on the mainland, such as that at Lerna, were destroyed by fire, and the houses built afterward were of a different type and more primitive. These new houses were long and narrow, only one story high, and apparently gable roofed. The entrance was at one end, and there was often a small compartment, which might be semicircular (apsidal), at the other. The new houses were evidently built by foreign invaders settling in the places they had destroyed. But some of the previous inhabitants may have survived as hewers of wood and drawers of water. There was continuity in the pottery, although new shapes and fabrics appeared alongside the old and sauce boats were no longer made. The site of the House of Tiles appears to have been reserved as sacred or unlucky ground, with a ring of large stones above its burnt ruins.

Either now or, possibly, somewhat later, there is evidence of a complete revolution in burial customs on the mainland. Family tombs were replaced by individual graves, usually lined and covered with slabs, such as the cist graves of the Early Bronze Age in the Cyclades. But these later graves on the mainland are often found inside the settlements, whereas before this burial outside the settlement area appears to have been the rule everywhere in the Aegean. The custom of "intramural" burial had lingered, however, in the hinterland of Anatolia, where houses were built in the style of those introduced to the mainland by the invaders who destroyed the mainland settlements *c.* 2200. Perhaps they came from there, but this remains uncertain.

About the same time as this invasion of the mainland, other people from Anatolia settled in some of the Cycladic islands, such as Syrus (Síros) and Ceos (Kéa) and in Euboea. On a rocky hilltop at Khalandriani in Syrus, they built an impressive fortified settlement that may have replaced an earlier native one that lay across a gorge to the south and had extensive cemeteries around it. It is clear from their pottery that these people who settled in the islands came from the coastal regions of western Anatolia, which had been comparatively civilized by then by sea contacts with Cilicia and Syria. Some of their pottery was made on a fast wheel, a useful invention that had originated many centuries earlier in Mesopotamia and began to be adopted on the mainland and in Crete at about this time.

The Middle Bronze Age on the mainland (c. 2000–1550). The mainland seems to have 0ndergone further disturbances at the beginning of the 2nd millennium. Settlements such as those at Lerna and Eutresis were destroyed again at this time. There is some slight evidence to suggest that the destroyers on this occasion came from the north. Small handmade pots of primitive appearance with possible relatives in the Balkans were found at Lerna in houses built immediately after this second destruction. Scraps of pottery of a different kind, with string-impressed decoration, at Eutresis may come from vases brought by the people who destroyed the settlement there. Corded ware such as this was made by the Kurgan people, who had spread from their original homeland in south Russia into the Balkans. Kurgan is the Russian word for a mound and refers to their custom of burying their dead in artificial mounds or barrows. Burial mounds similar to theirs are found in many parts of the mainland, notably in the western Peloponnese but also in Attica from this time onward. The Kurgan people were pastoralists and breeders of horses, and horse bones are first attested in mainland settlements at this time. Since the Kurgan people have a good claim to be regarded as Indo-Europeans, the first people of Indo-European speech, if not, as is widely believed, the first Greeks, may have begun to settle in southern Greece at this time. As often happened in ancient times, when pastoralists overran settled agricultural communities, these invaders appear to have adopted the material culture of those they had conquered. Longhouses of the kind introduced by the invaders a couple of centuries earlier remained in use, and the custom of burial inside the settlement, which is first attested for certain now, although it may have been practiced by the earlier invaders, continued alongside burial in mounds. There is some evidence for the existence of a new class of rulers who may have been the ancestors of the princes buried in the later shaft graves and tholos tombs of the mainland. An isolated mound at Elátia (Dhrakhmáni) in central Greece, covering the richly furnished grave of a princess, appears to be an early forerunner of these. A complex of buildings around a spacious yard dominated by a longhouse in the level immediately after the destruction at Lerna may have been the "palace" of the new ruler there.

The Kurgans

Middle Helladic culture

The development of civilization on the mainland continued to lag behind the Cyclades and Crete, and many centuries were to elapse before the art of writing spread from Crete to the mainland. But gold jewelry is found with some burials, and, by the time of the Mycenae shaft graves in the 16th century, a luxuriant style of native goldwork had been created. The people of the mainland remained in touch with the more civilized world of the Cyclades and Crete. Cretan pottery found its way to Lerna, and a fashion for decorating vases with designs in dark lustreless (mat) paint, which began on the mainland soon after the invasion, appears to have been adopted from the Cyclades. Much of the pottery on the mainland continued to be made on the fast wheel, and a class of gray burnished ware, already fashionable before the invasion, was now usually wheel made, with sharp angular shapes copied from those of metal vases. The polished gray surfaces of this "Minyan" ware (as it was named by Schliemann after the legendary inhabitants of Orchomenus in central Greece, where he first came upon it) look as if meant to imitate silver.

The Shaft Grave period on the mainland (c. 1600–1450). On the island of Cythera (Kíthira) between western Crete and the southern tip of the Peloponnese, a colony of Cretans appears to have replaced a settlement of people from the mainland toward the end of the 3rd millennium. In the 17th or 16th century, Cretan colonies were established at Triánda in Rhodes and at Miletus on the western coast of Anatolia. Later Greek legends seem to refer to colonies from Crete, if not from Knossos, in some of the Aegean islands. Much Cretan pottery found its way to the Cyclades and was imitated there; but, although the Cycladic people adopted fashions and ideas from Crete, they retained their own distinctive traditions. Cycladic imitations of the designs on pottery from Crete tend to be less refined but more lively than their Cretan models. Wall paintings found at Phylakopi in Melos may be the work of Cretan artists, but others of about the same date from the Akrotíri settlement on Thera, in spite of conventions and techniques learned from Crete, are very different in style and spirit from Cretan ones. Settlers from Crete, however, may have made their homes in Cycladic towns such as Phylakopi, and many of the islands might have been tributary to Knossos for a time.

The mainland, like the islands of the Aegean, came increasingly under the influencence of Cretan civilization, which seems to have extended to every aspect of life, including religion and dress. Its spread is vividly illustrated by the contents of the unplundered royal shaft graves at Mycenae.

These royal shaft graves belonged to leading, if not royal, families of the place that would eventually supplant Knossos as the chief centre of the Aegean. The burials in them seem to range over a period of 100 years or more, from shortly before *c.* 1600 to the beginning of the 15th century. The princes buried here were not Cretans, although it is possible that they paid tribute to Crete. The story of Theseus suggests that Athens was tributary to Knossos at one time, but there is no hint in legend of any such payment of tribute by the rulers of Mycenae. Two groups of shaft graves have been discovered at Mycenae in different parts of a large cemetery area. Each group was eventually surrounded by a circular enclosure wall. The circle designated B with the earliest burials lay outside the limits of the later Bronze Age defenses, but the other circle, A, enclosing the richest burials in six large graves, was deliberately incorporated within them. Schliemann excavated the graves of circle A in 1876, but it was not until 1951 that circle B was noted. The graves consisted of deep, rectangular shafts with stone-walled burial chambers at the bottom. After the bodies were placed in them, the chambers were roofed with timbers, and the shafts above filled again. Both men and women were buried in the graves, many of which contained several bodies. Sometimes the remains of earlier burials seem to have been pushed aside to make room for later ones; but, if so, the shafts must have been laboriously reopened to admit new burials. Large stone slabs with carvings in flat relief had been set above some of the graves. The carvings included spiral designs and pictures of the dead riding in their chariots to war or to the hunt. According to another interpretation, they are taking part in their own funeral games with chariot races as described in Homer. These tombstones provide the earliest evidence for chariots on the mainland. With the dead and, more especially, with those in the graves of circle A had been placed a fantastic array of gold and silver cups, jewelry, and dress ornaments. Golden diadems and elaborate hairpins decked the heads of women. Beads in necklaces were of amethyst, probably from Egypt, and amber from the Baltic. The men were buried with supplies of bronze weapons, including great slashing knives and spearheads and rapier-like swords of types in use in Crete at the time. Several of the swords had blades ornamented with reliefs; others had gold-plated hilts and pommels of polished stone or ivory. Over the faces of some of the men were gold portrait masks, which show them with beards and moustaches, unlike Bronze Age Cretans but curiously like Celts and Germans of Classical times.

Earliest evidence of chariots on mainland

Some bronze dagger blades were inlaid with remarkable pictures or designs in other metals, chiefly gold and silver and electrum (a mixture of gold and silver) in various shades. Black niello was used as a background for these pictures or to heighten the incised detail. The most famous of the shaft-grave daggers shows men armed with bows, spears, and great body shields, hunting a pride of lions; another has catlike animals chasing wild fowl among papyrus flowers beside a silver stream. This technique of "painting in metals" appears to have originated in Syria, although the workmanship and style of the pictures on the Mycenae daggers look Cretan. Whereas other daggers, however, and some metal cups with inlaid designs of this kind have been found on the mainland, none has yet been recovered in Crete. But many of the treasures from the shaft graves are imports from Crete, among them a faience ritual cup such as some from the palace at Knossos; others may

have been made by Cretan craftsmen working at Mycenae or by native craftsmen trained by Cretan masters. There were comparatively few clay vases in the rich graves of circle A but large numbers in those of the earlier circle B. While some of these clay vases were Cretan imports, most of them were locally made in traditional mat-painted or Minyan ware. A few, however, with designs in dark shiny paint on a light burnished surface, represent a new style that had developed on the mainland under Cretan influence. This Mycenaean pottery, very different from the mat-painted and Minyan wares it came to replace, seems to have originated in Messenia in the southwestern tip of the Peloponnese, whence it spread to other parts of the mainland and eventually as far north as Thessaly.

Mycenaean pottery

The use of Mycenaean pottery is the hallmark of the Mycenaean civilization of the Late Bronze Age. Other distinctive features of Mycenaean civilization that may have developed in Messenia under Cretan influence during the 16th century are the stone-built tholos tomb and the rock-cut chamber tomb. Cretan influence was so strong in Messenia at that time as to suggest the presence of communities of Cretan settlers there. But no positive evidence for the existence of Cretan settlements has been discovered. The tholos tomb (*tholos* is the Greek word for a round building), with high corbelled vault the shape of an old-fashioned beehive, is found in every part of the mainland to which Mycenaean civilization spread. But tholos tombs are most thickly concentrated in Messenia, and the oldest of them have been identified there. Some of the oldest of the Messenian tholos tombs resembled the early circular tombs of Crete in that they were built above ground and were used by families or clans for successive burials. But, in one important respect, they differed from the Cretan tombs. For, like all later tholos tombs, they were invariably covered with a mound. Burial mounds with many graves in them were common in Messenia throughout the Middle Bronze Age, continuing in use alongside tholos tombs in backward areas. The tholos tomb, which is in effect a circular tomb of the earlier Cretan type covered by a mound, appears to have been created therefore in Messenia either by settlers from Crete or more probably by natives of the region who had learned how to build round vaulted tombs in the Cretan manner.

Most tholos tombs on the mainland were built in a deep pit that was usually cut into the slope of a hillside. The top of the corbelled vault projected above the surface and was covered with a mound. The entrance to the circular chamber was roofed with huge slabs, and it was approached down a long passage (dromos) that was left open to the air. The most splendid of these tombs evidently belonged to kings or princes and may have been intended for only one occasion of burial. The few unplundered princely tholos tombs that have been excavated, notably one at Dhéndra (Dendra) near Mycenae, held several burials. Unless the tombs were reopened for successive burials, this would imply that wives and retainers were killed to accompany their lords and masters in death. The earliest of a group of three princely tholos tombs at Peristéria in Messenia was in use during the time of the Mycenae shaft graves. But the fashion for majestic stone-built tombs of this kind rapidly spread from Messenia to other parts of the Peloponnese, and at Mycenae they appear to have replaced shaft graves as the resting places of the local rulers about the middle of the 15th century. Several tholos tombs have been discovered at Mycenae, and the largest and most splendid, the so-called Treasury of Atreus and Tomb of Clytemnestra, appear to be among the latest. These great tombs, built of squared masonry and richly decorated, may have been constructed early in the 14th century, although arguments for a date as late as the 13th can be advanced. The mound that covered the Tomb of Clytemnestra overlay part of the earlier shaft grave circle B, and this might seem to imply a change of dynasty at Mycenae from rulers who buried in shaft graves to ones who used tholos tombs. But the later circle of graves A was respected and eventually incorporated within the defense walls when they were built in the 13th century.

The great tombs of Mycenae

While stone-built tholos tombs became the standard resting places for kings and princes in all parts of the mainland to which the Mycenaean civilization penetrated, the mass of the population changed from a custom of burial in single graves, whether in mounds or cemeteries or inside settlements, to the use of family vaults. In some regions, such as Messenia and the frontier area of Thessaly, families built small tholos tombs for themselves. But the most common type of Mycenaean family tomb was a rock-cut chamber with a dromos leading down to the entrance. The entrance was blocked with stones and the passage filled with earth after each burial. The rock-cut tomb may have been developed in Messenia during the 16th century under Cretan influence, like the tholos tomb. In the Knossos region of Crete, rock-cut tombs had been in use for communal burials for many centuries before this. Whatever its origin, the idea of family burial in rock-cut tombs soon spread to Mycenae and other parts of the mainland. Some rock-cut tombs in Messenia and elsewhere were carved in the shape of the beehive vaults of tholos tombs. A few large rock-cut tombs, including some of this shape, were used for royal or princely burials.

Period of the Early Palaces in Crete (c. 2200–1700). Crete does not seem to have been affected by the movements of people into the Cyclades and the mainland at the end of the 3rd millennium. But important changes were taking place there. Great palaces of a distinctive type built around large rectangular open courts seem to have been constructed within a comparatively short time at the leading centres of Knossos, Phaestus, and Mallia. The art of writing is first attested for certain in Crete at the beginning of this Palatial period. These developments in Crete appear to be the result of local evolution. It has been suggested that the models for the great palaces were Anatolian, but the Anatolian palaces, like that at Beycesultan, which have been compared with them, are different in plan and may be later in date. It is always possible, however, that immigrants from some other part of the civilized world of the Near East gave an impetus to these developments, if they did not impose their rule on Crete. Another suggestion is that refugees might have reached Crete from the mainland when it was invaded *c.* 2200. Seal impressions from the House of Tiles at Lerna, which was destroyed at that time, show designs reminiscent of those on Cretan seals of the period immediately before the rise of the great palaces. If the fashion for seals with designs of this kind became established in Crete only after the destruction of Lerna, mainland refugees might have been responsible for their adoption. But the dating of events on the mainland in relation to what was happening in Crete is not yet established with certainty, and the seals that made the impressions found at Lerna may have been Cretan ones or local imitations of them.

Earliest writing in Crete

At all events, Crete advanced rapidly along the path of civilization during the period of the Early Palaces, while the mainland relapsed into comparative barbarism. The art of the seal engraver made great strides in Crete at this time. Hard stones, such as jasper and rock crystal, began to be employed for some of the finer seals. A new and much favoured shape, which may have been adopted from Anatolia, was the signet with a stalk. Anatolian seals found their way to Crete and impressions of them have been identified in a great deposit of clay sealings from the early palace at Phaestus. Cretan seal designs now included elegant abstract patterns incorporating spirals and concentric circles neatly made with the drill, as well as lifelike pictures of animals, birds, and insects, together with mythical beasts, such as sphinxes and griffins adapted from Egyptian or Oriental models. Attractive hard stones, such as gabbro, were attacked by the Cretan vase makers, although they still made use of the softer chlorites and serpentines. Some of the fine stone vases from communal tombs in the Mesara region and at Mochlos may date from this period, rather than earlier, in the light of discoveries since 1950 in the early palace at Phaestus.

The fast potter's wheel began to come into use in Crete about the same time as in the Cyclades and on the mainland. Meanwhile, a revolution in the style of Cretan pottery was taking place. During the Early Bronze Age most of the finer vases everywhere in the Aegean area had been decorated with designs in dark rather shiny paint—

shades of red, brown, and black—on a light surface. But, toward the end of that period in Crete, there was a change to a "light-on-dark" style of decoration in which the vases were given an overall wash of the shiny paint previously used for decoration and designs were applied to this dark surface in white. This new "light-on-dark" fashion was also adopted, to some extent, in the Cyclades and on the mainland; but in Crete it was developed much further, and, from the beginning of the Palatial period, decoration in white was regularly supplemented with red to create a striking polychrome effect. This kind of pottery, which flourished in Crete throughout the time of the first palaces and later (*c.* 2200 to 1600), is known as Kamáres ware from a sacred cave of that name on Mt. Ida, where vases with fine polychrome decoration were recovered at the end of the 19th century. Most of the smaller vases in Crete, notably the drinking cups, now copy metal ones in their shapes and often in their molded or impressed decoration, and the exquisite "eggshell" ware, made in the workshops of the great palaces, with walls as thin as those of metal vases and shiny black surfaces adorned with abstract flowerlike designs in a combination of white, red, and orange, is among the finest pottery ever produced in Greek lands. The imitations in clay suggest that vessels of precious metal—gold and silver—were in general use in the palaces of Crete by this time. A silver, two-handled goblet of this period was recovered from a tomb at Gourniá in eastern Crete, and the treasure of silver cups, more than 150 in all, buried along with foreign lapis lazuli in temple foundations at Tod in Egypt in the time of Amenemhet II (1929–1895), may well be Cretan. Silver was more valuable than gold in Egypt at this time. It occurs in the Cyclades, and it was being mined before the end of the Bronze Age near Laurium in Attica on the mainland.

"Eggshell" ware

There were many contacts between Crete and the rest of the Levant during this period. Scarabs and stone vessels from Egypt reached Crete and were imitated there. Cretan Kamáres ware was exported to Cyprus, Syria, and Egypt, where it has been found in tombs and on town sites. Letters recovered from the ruins of the city of Mari on the Euphrates, destroyed by Hammurabi *c.* 1760, refer to objects of Cretan workmanship. It seems that Cretan metalworkers were already preeminent in the civilized world of the time. The daggers they made were of types ultimately derived from Syria, but they were exported to Cyprus in exchange, perhaps, for copper, although supplies existed in Crete. Westward, they may have reached Italy, where native copper daggers are of Cretan shapes and flint imitations of them seem to have been made. It was during this period that tin-bronze began to come into more general use in the Aegean, replacing copper or bronze made by adding arsenic, an effective process but dangerous for the craftsman who undertook it. Tin may have reached the Aegean first from Iran through Syria, although Etruria on the western coast of Italy was another possible source.

Burial in Crete was still normally in communal tombs, and many of the Early Bronze Age ones continued in use; but cemeteries of burials in storage jars are also attested at this time. No royal tombs of this period have been identified, however, and kings and queens may have been laid to rest, like their subjects, in the tombs of their clans. A large rectangular building with many rooms or compartments in the cemetery area just outside the city at Mallia might have been the tomb of the royal clan there. The local inhabitants plundered it during the 19th century, and its modern name—Chrysolakkos ("Gold Hole")—suggests what they found. A gold cup and jewelry, including elaborate earrings and pendants, acquired by the British Museum in 1892 and allegedly from a Mycenaean tomb on the island of Aegina near Athens, appears to be plunder from Chrysolakkos. French excavations there in the 1920s led to the recovery of similar jewelry, notably a gold dress-pin with flower head and a pendant in the form of a pair of bees (or wasps) facing each other over a disc, which may be meant for a honey cake. This pendant shows that the Cretan jewellers were masters of the art of hard soldering and could use it to fix wire (filigree) or minute globules of gold (granulation) to a background. But the jewelry from Chrysolakkos is difficult to date, and it may belong to a slightly later period, after the destruction of the early palaces at Knossos and Phaestus.

Aegina treasure

Life in the Cyclades seems to have continued much as it had in the Early Bronze Age. But, apart from signs scratched or painted on pottery from Phylakopi in Melos, there is little evidence of acquaintance with writing or the use of seals. Some time after the beginning of the period of the Early Palaces in Crete, Phylakopi was defended by a massive wall. Cretan Kamáres ware and Minyan ware from the mainland were imported to Phylakopi and doubtless reflect the trade in Melian obsidian, which may have been still in demand for cheap knives and razors, although metal ones were already in use in the Aegean area from the Early Bronze Age onward. Chamber tombs cut in the rock at Phylakopi appear to go back to this period, but burial in slab-lined cists continued elsewhere in the islands. At some point the fortified settlement at Khalandriani on Syrus was destroyed by fire and abandoned.

Period of the Later Palaces in Crete (c. 1700–1450). Various disasters occurred in Crete about the turn of the 18th and 17th centuries BC. The palaces at Knossos and Mallia were damaged, while that at Phaestus and a building that may have been the residence of a local ruler in a large settlement at Monastiráki west of Mt. Ida were destroyed by fire. What caused these destructions is uncertain. Accident, internal warfare, or foreign invasion are among possible agents. The damage at Knossos might have been caused by one of the earthquakes that afflict that area. It has been suggested that Crete was first conquered by Greeks at this period or by people from Anatolia speaking another Indo-European language called Luwian and related to Hittite. But there is no strong evidence for an invasion of Crete at this time.

Destruction of the palaces

The two or three centuries following these disasters were indeed the most flourishing of the Aegean Bronze Age, during which Cretan civilization reached its zenith. The palaces at Knossos, Phaestus, and Mallia were restored with greater splendour than before. From the dimensions of the new and entirely rebuilt palace at Phaestus, it has been possible to calculate the unit of length used by the Cretan architects: a foot only a fraction shorter than the standard English foot. In plan, the later palaces were basically the same as the earlier ones, with agglomerations of rooms clustering around long, rectangular central courts oriented roughly from north to south either for ritual or for catching the best of the winter sun. Many parts of these palaces were two or three stories high. A section on the eastern side of that at Knossos, built into a cutting in a steep slope below the level of the central court and housing the royal living quarters, may have had five stories. Large areas of the palaces, especially at Knossos, were reserved for cult, while other parts were occupied by storerooms for the produce brought as taxes in an age before the existence of money. Wide, paved squares flanked the palaces, and around them spread extensive towns, which, by this time, if not earlier, seem to have been unwalled. Cobbled streets with raised central paths of smooth squared blocks for the convenience of pedestrians ran through the towns. Surface water was carried away by covered drains, and skillfully jointed clay waterpipes were found in the palace at Knossos. The only settlement of this period that has been entirely excavated is a small town at Gourniá in eastern Crete. This was built on the slopes of a ridge overlooking the sea on top of which stood a little palace with a small open court in the centre and a public square beside it on the sheltered landward side. Down the ridge from the palace toward the sea was a small shrine facing the end of a path that led to it from the main street. Even in a small town such as that at Gourniá, many of the houses were evidently two stories high, and houses with three stories are depicted on faience inlays from the palace at Knossos assignable to the 17th century BC.

In palaces as well as houses, the lower parts of walls were still normally built of rough fieldstones held together with mud, the upper stories being continued in mud brick. But carefully squared and fitted blocks of limestone were employed for some important facades. Now, as earlier, walls were often laced together with a framework of timbers set vertically and horizontally in their faces and joined

by crossbeams running through them. There was also an extensive use of timber for columns and pillars and for the rafters supporting upper floors and roofs, which, it seems, were usually flat. Pictures of wooden columns show them with a characteristic downward taper, which may reflect an original custom of placing tree trunks upside down. The lower parts of the walls inside the palaces and great houses were often clothed with large slabs of attractively veined gypsum, a soft crystalline stone that outcrops in the region of Knossos and Phaestus. Gypsum was also much employed for pavements, but a hard lime plaster was more commonly used for coating walls and floors. Plastered walls were decorated with brightly coloured pictures, which may be an innovation of this period, since they are not yet attested for certain earlier in Crete.

Minoan frescoes

These pictures are described as frescoes because they were normally painted while the plaster was still damp. Lines impressed with string in the wet plaster helped to guide the artists. White, red brown, or blue were usually chosen as a background, while yellow and black were among other basic colours used. Many of these pictures, especially those from the palace at Knossos, were concerned with religion and show elaborately dressed goddesses, together with sacred dances and ceremonies, such as bull leaping, which may have had a religious or magical basis. But scenes such as a frieze of partridges and hoopoes adorning a room in what appears to have been an inn for strangers opposite the palace at Knossos look entirely secular. Monkeys, imported from Egypt, are depicted more than once, along with native wild goats and extraordinarily lifelike flowers, sometimes of identifiable species but often imaginary. The pictures ranged in scale from those with life-size figures, which might occupy most of the wall surface, to panels and friezes, including a class of miniatures with little figures of men and women two or three inches (five to seven centimetres) high. Parts of some wall pictures at Knossos were in relief, and plaster reliefs of this kind are occasionally found elsewhere in Crete. Floors and ceilings might also carry painted decoration.

Their wall paintings were probably the finest achievements of the Cretan artists, but only battered or fire-discoloured fragments of these have survived. The minor arts are better represented in the archaeological record. Now, if not earlier, hard rock crystal began to be used for making vases and seals, together with the volcanic glass, obsidian. A variety of this, flecked with spots of white pumice from Yialí (Glass Island), near Cos (Kos), was favoured for vases. Other fine stones imported for vase manufacture were Egyptian alabaster (calcite) and green and red marbles (antico rosso and lapis lacedaemonius) from the southern Peloponnese. Antique stone vases from Egypt might be adapted to local tastes by the addition of spouts and handles. Vessels with narrow necks were carved in two pieces that were afterward joined together, an example being a crystal libation vase from Zákros (Zakro) with the handle formed of crystal beads threaded on copper wire. A number of cult vases are carved with pictures in relief, including an octopus, a mountain shrine, a ritual dance of some kind, and games, such as bull leaping, wrestling, and boxing, which appear to have had magical or religious connotations. Soft stones, such as chlorite or serpentine, were used for making these vases, the surfaces of which were often coated with gold leaf to judge from scraps of it that have survived. This economical system of gilding was sometimes applied to seal stones, although

Rings and seals

solid gold and silver seals also occur. A class of gold signet rings has oval bezels engraved with ritual scenes that may be from the story of a goddess and her consort, with his death and resurrection. Seals of other shapes, in a wide range of attractive stones, display a variety of designs, including animals, such as lions, bulls, and wild boars. Sometimes a bull is being attacked by a lion, or a wild goat is escaping or standing at bay before a hound. Birds, fish, and butterflies also figure on these seals, and most of the designs appear to be entirely secular in character. But a class of gems crudely engraved with pictures of jars and leafy branches may have been rain-charms. Their skill in carving hard stones for seals and vases and their ability to model large plaster reliefs suggest that the Cretans might have sculptured stone reliefs and statues. But there is little evidence for Bronze Age sculpture in Crete, apart from a few small stone heads that may have come from statues with wooden bodies. Some bronze curls from the palace at Knossos appear to have adorned the head of a more than life-size wooden statue of a goddess. Figurines cast in solid bronze and mostly representing worshippers, both men and women, placed as votives in sanctuaries are often of great beauty. Statuettes of bull leapers and perhaps of gods and goddesses were made of imported ivory in several pieces cunningly joined together by pins and dowels. Faience manufacture was presumably learned from Egypt. Exquisite faience plaques of animals, along with statuettes of goddesses or priestesses and small vases of the same material, appear to be products of the palace workshops at Knossos.

The art of the potter, however, was in eclipse in Crete for a time, which may reflect the widespread use of gold and silver plate for tableware. Even cooking vessels were now being made of copper or bronze, including huge cauldrons in which a sheep or goat could be boiled whole. Among a variety of serviceable bronze tools were axes, adzes, and double-bladed axes such as those of earlier times. The sockets of these were improved toward the end of the period from a circular to an oval shape, which prevented twisting of the haft. New tools current by then included long bronze chisels and immense saws capable of slicing the gypsum required for paving and wall veneer, as well as for cutting timber. During this period, if not earlier, little trowel-like blades were replaced by leaf-shaped ones as razors. Helmets of copper or bronze are depicted on faience inlays from Knossos and stone relief vases, but plate armour is attested only from the end of the 15th century onward. For defense the Cretans of this time appear to have relied on huge rectangular or eight-shaped shields of bull's hide. (Homer's reference to the shield of Ajax "like a tower" preserves a memory of body shields of this kind.) Weapons included spears and daggers, as well as rapiers with long slender blades and short tangs for affixing wooden hilts. Massive pommels of attractive stones, such as rock crystal, or of gold-plated wood or ivory helped to balance the blades. Toward the end of the period, swords are found with strong, flanged hilts and short blades adapted for cutting as well as thrusting strokes.

Signs scratched or painted on clay vases, not only in Crete but on the mainland and in the islands, from about the middle of the 3rd millennium onward may reflect acquaintance with writing among the peoples of the Aegean area. But the first positive evidence for the use of writing in the Aegean is found in Crete at the beginning of the Palatial period—*c.* 2200 or somewhat later. This earliest Cretan writing is known as pictographic or hieroglyphic because its signs are pictures of animals or things in the manner of Egyptian hieroglyphs. But only a few of the signs resemble Egyptian ones, and the system appears to be of Cretan origin, even if it was inspired by Egypt or Syria. During the period of the Early Palaces and while the Cretan hieroglyphic script was still in use, a simplified linear script was being scratched on clay tablets at Phaestus. A more evolved script with linear signs of this kind is attested in various parts of Crete during the period of the Later Palaces and is known as Linear A, to distinguish it from the variety of script (Linear B) current both in Crete and on the mainland from the end of the 15th century onward (see below). Most of what has survived of Aegean Bronze Age writing is on clay tablets of the kind used in Syria and Mesopotamia in early times. But ink was used to write Linear A inscriptions around the insides of two clay cups from Knossos, and the bulk of what was written in the Aegean during the Bronze Age may have been in ink on some kind of paper, made from papyrus as in Egypt or from palm leaves, as later Greek tradition hints. Palm trees grew in Crete during the Bronze Age and some of the later (Linear B) clay tablets there and on the mainland seem to have been made in the shape of palm leaves.

Linear A and Linear B

THE DECLINE OF THE EARLY AEGEAN CIVILIZATIONS

The eruption of Thera (c. 1500) and the conquest of Crete (c. 1450). Cretan civilization reached its highest peak in

the 16th century. An important change of fashion that began in Crete then was the abandonment of the "light-on-dark" style of vase decoration of Kamáres tradition in favour of a return to "dark on light." The new-style Cretan pottery, with attractive designs of plants and flowers in shiny black or brown paint, was soon to inspire the development of Mycenaean pottery on the mainland. But this flourishing period in Crete ended in a series of disasters. About 1500 the volcano on the island of Thera, long, it seems, quiescent, erupted to bury the settlements there under many feet of pumice ash. The story of Atlantis, if Plato did not invent it, may reflect some Egyptian record of this eruption, one of the most stupendous of historical times. Knossos was shattered by a succession of earthquakes that preceded or accompanied the eruption, while great waves resulting from it appear to have damaged settlements along the northern coast of Crete. Later Greek traditions, such as the story of Deucalion's flood, may enshrine a memory of similar waves that swept the coasts of the mainland at this time. Some Cretan settlements might have been wrecked by blast from the eruption, although Thera lies about 70 miles (110 kilometres) away from Crete. Whatever the damage caused, it appears to have been soon repaired. Similarly, in the Cyclades there are few signs of any gap in occupation as a result of the eruption. The so-called Second City at Phylakopi in Melos was totally destroyed, but the survivors appear to have set to work at once to build the Third City on its ruins. The settlements on Thera, however, lay buried deep in pumice, and the largest of these, at Akrotíri, opened by excavations since 1967, offers a unique picture of a Bronze Age town, the walls of the houses standing in places two stories high, with paintings miraculously preserved on them, and the floors with storage jars and other objects as they were left when the inhabitants escaped from the eruption or from the earthquake that is thought to have preceded it.

After the eruption, Crete appears to have enjoyed comparative prosperity for a time, while the influence of Cretan civilization continued to spread on the mainland. Alongside vases with plant and flower designs, the Cretan potters began to make ones decorated in an attractive marine style, with octopuses and other sea creatures. The marine style may have originated at Knossos, but vases with this type of decoration, many of them of a ritual character, were exported all over Crete, as well as to the Cyclades and the mainland. About the middle of the 15th century, however, a generation or so after the eruption of Thera, most of the important sites in central and southern Crete were destroyed by fire. Whether this horizon of fire destruction can be traced in western Crete is uncertain because of the lack of excavation there. Destruction was not confined to palaces and towns but extended to country houses, farms, and rural shrines. Many settlements were never inhabited again, such as that at Mochlos, where the excavators found the remains of numbers of people who had perished in the destruction. The site of the destroyed town at Gourniá was eventually occupied by a scatter of houses, but the palace there was not rebuilt. The large palaces at Mallia and Zákros were also destroyed by fire and afterward abandoned.

The marine style

A new social order. The fact that palaces and country houses, centres of landed estates, were not rebuilt suggests a total overthrow of the existing social order. A number of magnificent stone ritual vases and bronze tools have been recovered from the ruins of the palace at Zákros in excavations since 1962, but virtually no gold or silver objects were found. Indeed, it looks as if the palaces and houses everywhere in Crete had been ransacked before they were destroyed. Alone of the four great palaces, that at Knossos may have escaped serious damage at this time, but parts of the city there were wasted by fire. In the early days of Cretan exploration, it was taken for granted that these destructions were the result of war. Since the 1930s, however, it has been suggested that they were, in some way, caused by the eruption of Thera. But the eruption began and ended a generation or more before this horizon of destruction, while conquerors are found in Crete immediately after it. The destruction appears to have been their work.

They evidently came from the mainland and made their capital at Knossos, but they seem to have established another centre of power at Phaestus; and legend hints at a third based on Kydonia (modern Khaniá) in western Crete. Vases of shapes already popular on the mainland, such as drinking cups with high feet, became fashionable at Knossos after the conquest and eventually spread to other parts of the island. A rather stiff, formal "Palace Style" of vase decoration, using motifs derived from the earlier plant and marine styles, may reflect an adaptation of Cretan fashions to cruder mainland tastes. The old clan tombs went out of use in the Knossos region and were replaced by rock-cut tombs of the type evolved under Cretan influence on the mainland. Some of these contain the burials of warriors and their families, accompanied by rich assortments of weapons and jewelry, resembling the Mycenae shaft graves and the mainland tholos tombs. There was a cemetery of similar rock-cut tombs, with richly furnished burials (the Tombe dei Nobili), at Phaestus. Tholos tombs sunk in the ground and covered by mounds appear to have been introduced to Crete from the mainland now. One, on the Kefála ridge north of the palace at Knossos, may have been built soon after the conquest.

A decline in taste and artistic imagination becomes marked during the period after it throughout the Aegean area. But skills and techniques remain at a high level, and the fabric and firing of pottery improve both in Crete and on the mainland. Great advances were made in military equipment, to judge from the formidable and often elegant and richly ornamented bronze weapons from the warrior graves of the time. Short, strong swords had blades adapted for cutting as well as thrusting. Their hilts were sometimes sheathed in gold, and pommels were made of fine stones or ivory, as in the past. Warriors were armed with large thrusting spears as well as throwing spears or javelins. Bronze arrowheads, virtually unknown in Crete before the conquest, are of types introduced from the east or copied from the flint and obsidian arrowheads long current on the mainland. Concial helmets of felt or leather might be cased in bronze, such as one from a warrior grave of *c.* 1400 at Knossos, which was surmounted by a knob for holding a plume and had separate cheekpieces. An old mainland custom of plating helmets with the tusks of wild boar is first attested in Crete during this period. Ivory plaques from the mainland show men wearing such helmets, and Homer describes one in detail. One placed in a tomb at Dhéndra, near Mycenae, *c.* 1400 had separate bronze cheekpieces, and the warrior to whom it belonged also had a pair of bronze greaves (leg guards) and a corslet of bronze plates cunningly fitted together. Similar corslets are depicted on Linear B clay tablets from Knossos as part of the issue to chariot warriors. Plate armour of this kind may have been developed in connection with chariot warfare for which the old body shield was impractical. Small, round shields also began to come into use in the Aegean area at about this time. Some Aegean ships by now had railed platforms at prow and stern joined by a gangway down the length above the heads of the rowers. Before the end of the Bronze Age, ships of this type had been developed as warships with rams.

Advances in weaponry

Events on the mainland. While there are many signs of mainland influence in Crete in the period after it, the conquest may have helped to spread Cretan fashions and techniques on the mainland through the medium of captive artisans sold as slaves. The earliest wall paintings on the mainland appear to date from this time and are thoroughly Cretan in style. But the Mycenaean civilization of the mainland remained very different from that of Crete. Mycenaean pottery is distinguishable from Cretan, and religious customs, such as worship in caves or hilltop sanctuaries, which continued in Crete, do not appear to have taken root on the mainland. The standard mainland palace of this period, although built with Cretan techniques and adorned with wall paintings by Cretan artists or in the Cretan manner, differed from the traditional Cretan palace centred around a large, rectangular court. The focal point of a mainland palace, such as that at Pylos (Pílos) in Messenia, was a great hall with the roof supported on four pillars and a vast circular central hearth. The hall

Mainland palaces

was entered through an anteroom with a columned porch beyond it. This complex appears to be an adaptation of the type of longhouse found on the mainland since the end of the 3rd millennium. The descriptions of palaces in Homer are evidently based on memories of palaces such as these, and what Homer calls the megaron corresponds to the great hall.

Conquest of the Cyclades. People from the mainland overran the Cyclades when Crete was conquered or a few years later. Settlements such as that at Ayía Iríni on Ceos (Kéa) were destroyed by fire, and the old Cretan colony on Cythera was abandoned. The so-called Third City of Phylakopi on Melos suffered damage, and, immediately afterward, a small palace of mainland type was built there. Similar palaces may have been built in Crete after the conquest. The largest house of this period at Gourniá is very similar in plan to the palace at Phylakopi. Massive foundations at Ayía Triádha near Phaestus may have belonged to the hall of a palace of orthodox mainland type. But the great palace at Knossos retained its traditional Cretan form, in spite of many alterations that were made to it. The paintings on its walls in the period after the conquest were still largely of a religious nature and included scenes of bull leaping and friezes of people taking part in processions or ritual. These friezes were inspired by Egyptian models, although their monumental character may have appealed to mainland tastes. The rulers of the mainland also adorned their palaces with religious pictures of this kind, but they appear to have indulged a passion for scenes of warfare and hunting to a degree that is not attested at Knossos.

Linear A was the script in use in Crete at the time of the conquest. But this had been replaced by Linear B by the end of the 14th century. Linear B clay tablets have been recovered from Knossos and from several of the palaces on the mainland. The people of the Cyclades may have been acquainted with the Cretan Linear A script, to judge from signs of it on local vases from settlements on Thera buried by the eruption of *c.* 1500, but it is doubtful if a version of this script was used on the mainland. If it was, the people of the mainland might have adapted it to form the Linear B script for writing their own language before they conquered Crete. Alternatively, Linear B may have been evolved afterward at Knossos. The Linear B tablets consist mostly of short lists or inventories, and picture signs (ideograms) often give an indication of what is listed on them. Thus, some tablets from Knossos, with a chariot, horse's head, and corslet, evidently record equipment issued to warriors. Ever since the first Linear B tablets were discovered at Knossos in 1900, many people have thought that their language might be Greek, and it was deciphered as Greek by Michael Ventris in 1952. His decipherment has been widely accepted, but, with short inscriptions of the kind available, it is difficult to verify. Those who doubt his decipherment include some who agree that the language of the Linear B tablets is Greek, whereas others think it is a non-Greek language spoken by the conquerors of Crete or a form of the old Cretan language that they adopted for official purposes at Knossos and on the mainland.

The fall of Knossos and the Mycenaean empire. The new rulers of Crete evidently thrived for several generations after the conquest. Great tombs were built and richly furnished. The unplundered burial of a princess of this period has been found at Arkhanes near Knossos. Then, suddenly, this comparative prosperity was brought to an end. The palace at Knossos was destroyed by fire and never rebuilt. Since 1960 the date of this final destruction of the Knossian palace has been disputed with a range of opinion from *c.* 1400 to 1150, but it probably took place in the 14th century, when there is evidence for destruction at other Cretan sites, such as Mallia, and when burial in the warrior graves of the Knossos region and at Phaestus ceases. The elimination of the warrior aristocracy that had flourished in Crete since *c.* 1450 was surely the work of enemies, possibly from the mainland. But the mainland also appears to have undergone changes for the worse about this time. The era of great princely tholos tombs on the mainland seems to come to an end now or not much later. From the beginning of the 14th century, the arts of wall painting, seal engraving, and vase decoration everywhere show symptoms of stagnation and decline. The pottery of the 13th century is remarkably standardized and homogeneous throughout the Aegean area. After the fall of Knossos, the chief centre of power in the Aegean appears to be Mycenae, which may have exercised control over an empire including Crete and the islands. From *c.* 1400 or earlier, the records of the Hittite Empire in Anatolia mention a state called Ahhiyawa, which may be some Mycenaean dependency, if not an empire ruled by Mycenae. Ahhiyawa could be the same word as Achaioi (Achaeans), the name given to the Greeks in Homer.

Final destruction of the palace at Knossos

The end of the Bronze Age in the Aegean (c. 1200–1000). There is no evidence that any palaces existed in Crete after the final destruction of that at Knossos; but, on the mainland, palaces continued to flourish at centres such as Mycenae, Tiryns, and Pylos. In the 13th century massive defense walls were built at many of these mainland centres, including Mycenae and Tiryns and Athens, where the acropolis was fortified. At Athens and Mycenae underground passages led to springs concealed beyond the walls. The Lion Gate at Mycenae dates from this time, although the limestone relief of confronting lions above it may have been carved earlier. A wall appears to have been constructed during this period across the Isthmus of Corinth against enemies from the north.

Toward the end of the 13th century, the chief centres of the Peloponnese were destroyed by fire, and their palaces were never rebuilt. Centres farther north, such as Thebes and Gla, may have been destroyed some time before this. The site of Pylos was afterward deserted, but occupation of a kind continued at Mycenae. This destruction of Mycenaean centres on the mainland seems to herald a collapse of civilized life throughout the Aegean area. Wall painting, gem engraving, and writing vanish from the scene; but a more efficient type of slashing sword appears in the Aegean about this time, and the skill to make complicated military equipment, such as chariots and warships with rams, was not lost, while eventually iron replaced bronze for weapons and tools, as elsewhere in the Near East. After the destruction of the Mycenaean centres in the Peloponnese, there is evidence for depopulation there and for a great exodus of the former inhabitants, some westward and southward to Cephallenia and Crete, others eastward to the Cyclades and as far as Cyprus and Cilicia, where their settlements have been noted. An important cemetery of this period has been excavated at Peráti on the eastern coast of Attica.

Introduction of iron

There has been much speculation as to the causes of these disasters and upheavals. Revolutions or civil wars and vast piratical raids by land or sea have been blamed for the destruction of the Mycenaean centres. The dispersal of refugees from the Peloponnese has been explained in terms of climatic changes leading to drought and famine there. But the evidence for such changes is doubtful, and the refugee movements look very much like flight before the onslaught of invaders.

Invasion from the north. There are indeed signs that barbarous peoples from the north had begun to enter the Peloponnese by the 12th century. Traces of them are slight and controversial, because they adopted elements of the Mycenaean civilization, and most of their pottery was made on the fast wheel in a version of Mycenaean style and, at first, presumably by Mycenaean craftsmen whom they had enslaved. Burial in single graves, current on the mainland before the rise of the Mycenaean civilization, had lingered in regions beyond its frontiers, such as Epirus and northern Thessaly. Now, single graves, which may be those of invaders, are found in the Peloponnese and elsewhere on the mainland alongside family tombs of Mycenaean tradition. Pairs of long dress pins, with burials of women, are among northern features that appear in the Mycenaean area at this time. These pins, worn one on each shoulder, suggest the introduction of a new type of garment, different from anything Mycenaean and akin to the simple blanket-like peplos still retained by the womenfolk of conservative Dorian Greeks in Classical times.

The coming of the Greeks. The Classical Greek dialects

fall into an eastern and a western group, which might reflect two successive movements of people of Greek speech into Greece. According to tradition, the Dorians, or West Greeks, only entered the Peloponnese in the 11th century. It is therefore difficult, although not impossible, to regard them as the destroyers of the Mycenaean centres there *c.* 1200. As to when the bearers of the East Greek dialects first appeared in Greece, there is a wide range of opinion. At one extreme is the idea that the Neolithic inhabitants of the Aegean already spoke a form of Greek. According to another view, both groups of Greek dialects were actually formed on Greek soil early in the Bronze Age, the East Greeks being the creators of the Mycenaean civilization, while the West Greeks remained in primitive conditions beyond its frontiers until the 12th century or later. One widely held view brings the ancestors of those who spoke the East Greek dialects from Anatolia or the north in the centuries around 2000, when there is evidence of successive invasions on the mainland. At the other extreme is the belief that Greek was not spoken in the Mycenaean area until *c.* 1200, when the bearers of the East Greek dialects came as the destroyers of the Mycenaean centres in the Peloponnese and adopted a barbarized version of the Mycenaean civilization as their own. This view, however, is naturally difficult to reconcile with a belief that the language of the Linear B tablets is Greek.

Theories of the origin of Greek dialects

The people of the Bronze Age Aegean. *Physical appearance.* The inhabitants of the Aegean area in the Bronze Age may have been much like many people in the Mediterranean basin today, short and slight of build with dark hair and eyes and sallow complexions. Skeletons show that the population of the Aegean was already mixed by Neolithic times, and various facial types, some with delicate features and pointed noses, others pug-nosed, almost negroid, are depicted in wall paintings from the 16th century. But men and women are always represented with black hair, and the presence of fair-haired people is not attested in the Aegean until later Greek times. Some very tall men buried in the Mycenae shaft graves may be descendants of invaders who entered the mainland at the end of the 3rd millennium. A few skeletons from the single graves that appear on the mainland at the very end of the Bronze Age suggest the presence of new people from the north.

Dress. Clay figurines of *c.* 2000 from Crete show men wearing a narrow codpiece with a belt or loincloth and bare above the waist. This was to remain the basic fashion for Cretan men throughout the Bronze Age. Cretan women wore short-sleeved jackets, which left the breasts bare, and ankle-length flounced skirts, although shorter skirts to just below the knees are also attested. Marble figurines of men from the Cyclades assigned to the Early Bronze Age have belts and narrow codpieces such as the Cretans. There is little evidence for dress on the mainland until the time of the Mycenae shaft graves in the 16th century. A considerable variety of dress is represented from that time onward throughout the Aegean area, but it is difficult to recognize fashions peculiar to the mainland. Tasselled shorts worn by men on the Mycenae lion-hunt dagger are also attested in Crete at the time. Aegean envoys painted on the walls of the tomb of Rekhmire, vizier of the Egyptian pharaoh Thutmose III (ruled 1504–1450), at first wore large codpieces of a type fashionable in Crete in the Shaft Grave period, but these were later changed into kilts without codpieces as worn by men in paintings at Knossos after the mainland conquest of Crete, *c.* 1450. This might reflect mainland envoys coming to Egypt after the conquest and wearing a different type of dress from the Cretan. But kilts of this kind appear to be represented in Crete both before and after the conquest. Curious scaly cloaks and long single-piece robes are among a variety of ritual garments in Crete. Linen was known in Crete by the beginning of the Bronze Age and fragments of it were recovered from the Mycenae shaft graves. But in Crete, at any rate, clothes were mostly, it seems, made of wool, and wall paintings show them woven with colourful and intricate designs, including pictures of animals and birds and even musical instruments. One of the dyes used was purple crushed from murex shells. Cretan men wore knee boots and sandals with upturned toes. Men with leggings or greaves are represented in wall paintings on the mainland. Caps of various kinds appear on the heads of men and high, pointed hats and tiaras on those of women and of gods and goddesses or their priests. Clay figurines of the Early Palace period show Cretan women with elaborate hair arrangements. Women put jewelry in their hair, including strings of beads. Necklaces, earrings, bracelets, and armlets were displayed by men as well as women. Seal stones were carried on strings around the neck or on the wrist. Cretan men normally left their hair long but were clean shaven. Beards and moustaches are attested on the mainland in the Shaft Grave period and later.

Society. In the 3rd millennium the whole Aegean area appears to have been divided into small units ruled by chiefs with "palaces" not much larger than houses. But the Early Palaces in Crete suggest larger kingdoms there by *c.* 2000, with a stratified society including nobles and probably classes of scribes and free artisans, as well as slaves. In the period of the Later Palaces, if not earlier, every sizable settlement in Crete was dominated by a large house, or palace, evidently belonging to the chief family of the place. These families may have formed a nobility owing allegiance to the kings of the great palaces at Knossos, Phaestus, Mallia, and Zákros. The kings may have shared power with councils of nobles, if not with popular assemblies, but this is speculative and based on doubtful analogies with early Sumeria. Primitive customs of the kind loosely called matriarchal, such as inheritance through the female line, may have survived in Crete, and women evidently played an important part in religious cult there. The Mycenaean society of the mainland appears to have had an aristocracy of conquest. Toward the end of the Bronze Age, the evidence suggests the formation of larger states extracting tribute from local rulers and eventually perhaps suppressing them to exercise direct control.

Social organization

Trade. Foreign manufactures reaching the Aegean and especially Crete during the Bronze Age included Cypriot pottery, Mesopotamian and other Oriental cylinder seals, and Egyptian stone vases, ivories, and scarabs, while Cretan and eventually Mycenaean pottery is found in Egypt and elsewhere in the Levant. But the bulk of what was traded evidently consisted of perishable fabrics and foodstuffs and, above all, raw materials. Obsidian from Melos reached other parts of the Aegean from Upper Paleolithic times onward. Imported tin became a necessity during the Bronze Age, and copper was brought from Cyprus to supplement native sources. Massive copper or bronze ingots of an international type are found in Crete and elsewhere in the Aegean from the 15th century onward. Ostrich eggs and stone for making vases were among imports to Crete from Egypt, and ivory came from there or from Syria. Amber from the Baltic reached the mainland in some quantity during the Shaft Grave period and later but is rarely found in Crete. Exports from the Aegean may have included woollen goods, olive oil, and timber, as well as silver. In Crete, at any rate, foreign trade may have been largely under palace control, but a class of private merchants engaged in overseas commerce no doubt existed in the Aegean.

Transport. Ships with a mast and square sail in addition to oars or paddles are attested in the Aegean from the Early Bronze Age. On land, goods were no doubt transported by pack animals or on poles slung between bearers, a principle also adopted for passenger chairs, of which there are clay models. A model of a four-wheeled cart from Crete is datable to *c.* 2000 or earlier. The wheels of such carts were evidently solid, and they were no doubt drawn by oxen, although horses may have been ridden in Crete by then, as they seem to be depicted on early Cretan seals. These horses could have come from the east, a different breed being introduced into the mainland from the north at the beginning of the Middle Bronze Age *c.* 2000. The light spoke-wheeled chariot drawn by horses appears to have developed in Syria or northern Mesopotamia early in the 2nd millennium but spread rapidly throughout the Near East, owing to its usefulness in war. Chariots are depicted on tombstones of the Mycenae shaft graves and on Cretan seals before the time of the mainland conquest

c. 1450. Apart from warfare, they were used in the Aegean for hunting and probably for travel. During the latter part of the Bronze Age, terraces were built to support roads wide enough for wheeled vehicles both in Crete and on the mainland. Such roads were carried across streams on bridges, examples of which have survived in the region of Mycenae.

Warfare. Short daggers of types derived from Syria were in use in the Aegean during the Bronze Age. Long rapiers, evolved from those in Crete, are found on the mainland by the time of the Mycenae shaft graves in the 16th century BC. Short swords adapted for cutting as well as thrusting begin to appear in the following century and may have been developed in connection with chariot warfare. Bronze armour and small, round shields more serviceable in chariots replace the old Cretan body shields at approximately the same time. Bows and slings were probably used everywhere in the Aegean area, but, whereas arrowheads of flint and obsidian are found on the mainland, they are virtually unknown in Crete, where arrows may have been tipped with bone or wood until the appearance of bronze arrowheads in the 15th century.

Development of fortified towns

Settlements on the mainland and in the Cyclades were defended by walls from the Early Bronze Age onward, and the town at Mallia in Crete appears to have been protected by a wall during the period of the Early Palaces; but, by the time of the Later Palaces, Cretan towns may have been unwalled. Faience inlays of the 17th century from Knossos, however, seem to show an attack on a walled town such as that depicted on a silver-relief vase from the Mycenae shaft graves. The town, in each case, appears to be in the Aegean but not necessarily in Crete. Methods of warfare had become highly developed by the end of the Bronze Age, with improved weapons, complex and well-designed fortifications, extensive use of chariots, and warships with rams.

Religion. Little is known about religion in the Cyclades and on the mainland before the period when they came under strong Cretan influence. An open-air sanctuary filled with marble figurines on the island of Kéros (Káros) is assignable to the Early Bronze Age. In Crete during the Early Palace period, there were many open-air sanctuaries on the tops of hills and mountains. Some of these had small shrines in them, and shrines with one or more rooms and benches for offerings and cult statues are found in the countryside and in the towns in Crete. Parts of the palaces and of large houses there were also set apart for cult. Shrines not unlike Cretan ones existed in settlements in the Cyclades and on the mainland in the Late Bronze Age; but hilltop sanctuaries are not well attested there, and most of those in Crete appear to go out of use after the mainland conquest, *c.* 1450. Caves were also used as sanctuaries in Crete, and cults in some of these continued until the end of the Bronze Age and later.

The chief deity everywhere in the Aegean in the Bronze Age was evidently a goddess. Perhaps there were several goddesses with different names and attributes. Clay statues of goddesses have been found in shrines in Crete dating from the 15th century. Shrines with clay goddess figures have also come to light at Ayía Iríni in Ceos and at Mycenae. But most cult statues were probably made of wood everywhere in the Aegean throughout the Bronze Age. There is evidence for the worship of a male god, who may have been a counterpart of gods such as the resurrected Osiris in Egypt. Much of Aegean Bronze Age religion survived into later Greek times, and the idea of an Elysion, or Isles of the Blessed, very different from the gloomy Hades of Homer, appears to be a legacy from Bronze Age Crete. (M.S.F.H.)

Ancient Greek civilization

The special role of Greece in the history of Western civilization is easier to demonstrate than to explain. The land itself offered no particular advantages such as far earlier had determined the importance of the great valley civilizations of the Nile, Mesopotamia, or the Indus. It plays little part in the history of early man. Its Bronze Age history has something to contribute to the history of civilization but is trivial beside the great contemporary empires of the East and Egypt. Something that may help explain why Greece emerged as a great civilization is the happy combination of geography and temperament: a geographical location that permitted it to receive the benefits of the skills and resources of older civilizations and contemporary civilizations to the north, east, and south, and yet that was not especially vulnerable to attack by them; and a spirit of inquiry, bred, perhaps, in far earlier generations of nomadic life, which was manifested by the Greeks' exploration, travel, and colonization and by a speculation about man and his world, which soon led to a sound basis for philosophical, political, and even scientific theory. Another important factor was a language that developed into the most subtle instrument for the expression of thought, science, and emotion that the West has known.

Achievements of Greek civilization

The achievements and legacy of Greek civilization can be assessed under many different heads, and yet, the most significant contribution was probably an attitude of mind and an approach to political life rather than works of art and literature. The autocratic dynasties of the East and Egypt with their royal and priestly hierarchies could provide no model for politics in Greece. Instead, the small kingdoms progressed from systems of rule by families or dictators (in the modern sense) to rule by elected assemblies and councils and to recognition of the virtues of democracy, as well as of its practical shortcomings. The practice of government came to be systematically analyzed for the first time, and the systems of different contemporary states compared. The problems of the slaves and the poor could be recognized, if not solved. In the fine arts, the technical skills of the Greeks were mainly learned from foreigners, except perhaps for the art of mural painting. It was in terms of conception and precision that Greek artists outstripped their teachers and set standards of artistic expression and craftsmanship that seem almost superhuman. Their special contribution to narrative art lay in their genius for giving expression to subtleties of mood and feeling while seemingly confining themselves to mere reportage. Their graphic art exhibits similar qualities, expressing a perfect compromise between, on the one hand, a sense of pattern and proportion, and, on the other, shrewd observations of man and the world around him.

Ancient Greece was the most literate of all ancient cultures—witness the profusion of public and private inscriptions, from state decrees to graffiti on walls. Some of the earliest uses of the newly learned alphabetic script were for epic poetry; for law codes; and for lyric poetry, private and public, on themes of state, war, humour, and love. Eventually, there emerged choral and dramatic treatments of the traditional myths, done in a spirit of moral speculation, religious affirmation, or sheer irreverence. The great prose histories are models of disinterested inquiry and of a critical concern for recording man's motives as well as his behaviour.

The attempt of Greek philosophy to provide a theory of the universe to replace the cosmologies of myth eventually led to practical scientific discoveries. And, by exploring man's processes of thought and argument, principles of description and logic were defined that were both an expression of and an influence upon the practice of political and personal morality. The search for principles was in fact carried so far that, too often, theory was preferred to the more mundane process of observation, so that progress in the natural sciences and medicine was slow compared to achievements in other fields.

That man is the measure of all things and mankind his special study may appear today the most remarkable tenets of ancient Greece, especially in light of the obsessional religions, oppressive governments, and conventionalized arts of other early societies. This is, on the whole, a just appreciation of the Greek achievement, but in the past it has led to an unreal, not to say sentimental, idealizing of Greek standards of virtue, wisdom, and beauty. More recent studies have dwelt more objectively on the Greeks as one society out of many in antiquity, whose debts to others were at times considerable and whose success must be measured against the not inconsiderable achievements of other peoples. But it remains true that no other ancient

people exhibited such a wide range of genius or left such a vigorous legacy. The heroic quality of Greek history—the stand against Persia, the exploits of Alexander, expressed so proudly by their own historians and poets—is, if anything, enhanced by this new objectivity, by the understanding that the Greeks were subject to the same fears, hopes, motives of power, self-interest, love, duty, and cruelty that have beset men throughout history.

Influence of geography on Greek life

The Greek landscape is that of the miniaturist—with rapid but never drastic transitions from valleys to hills and, for such a small country, with so many geographically distinct enclaves, including its many islands, that it is totally unlike the broader tracts of plains and mountains that are characteristic of most other Mediterranean lands. This simple geographic factor explains the separate development of the Greek city-states, which were bound by a common language and interest rather than by any sense of belonging to a distinct geographical entity—until, that is, a sense of national unity was imposed on them from outside. Greece's long coastline and the difficulties of land communication meant that seafaring always played an important part in Greek life and politics and probably contributed to the Greeks' readiness to explore the farther Mediterranean shores. Unlike Asia Minor, or even Italy, Greece is a country of sea routes, not roads.

The provinces of Greece represent geographic, racial, or dialect unities rather than political unities. In the north, Macedonia, a peripheral Greek area until the age of Philip and Alexander, marked the end of the main routes from Europe. To the south, the country is split lengthwise by the Pindus Range, with Epirus and Aetolia to the west, facing the Adriatic, and the broad Thessalian plain sloping to the Aegean. The mountain range is difficult to cross at any time and almost impossible in winter. East central Greece, east of Mt. Parnassus and Delphi, forms a separate enclave with the plain of Boeotia, the hillier country of Attica with its thinner soil, and the long and rather bare island of Euboea flanking them. The Corinthian Gulf nips the peninsula, making the isthmus to the Peloponnese a natural focus of routes and power. The isthmus was dominated by Corinth, which built a channel (*diolkos*) for dragging ships overland. From the east, the seaward approaches were controlled by Athens and Aegina, while Megara, north of the isthmus, must have owed much of its wealth to its position beside the only north–south route overland. The mountains of south Greece, the Peloponnese, were equally effective as dividers of the land. There are only two plains of any importance, serving Argos and Sparta, and the coastline is generally inhospitable. The jutting masses of Euboea, Attica, and Argolis are continued southeast in a broad scatter of islands reaching across to the southwest corner of Asia Minor. These are the Cyclades, with Dorian Rhodes and the Dodecanese at the eastern end. South lies the great island of Crete, halfway from Athens to Africa, geographically the southern border of the Aegean world.

The west coast of Asia Minor is quite unlike Greece, with broad river valleys, massive mountain ranges running inland to the east, deep gulfs, and large offshore islands. The Greeks, whose communications ran more readily by sea than by land, naturally came to regard the coastal areas and islands of the east Aegean as theirs to exploit, which they did through much of the period, coming regularly into conflict with the mainland powers (Lydia, Persia), which could not brook foreigners on their flanks.

The Greek climate

The climate of Greece is temperate, with long dry summers favouring the cultivation of vine and olive, and wet winters, especially in the western areas, but not generally very harsh. Barely one-fifth of the land area is arable, and, in parts of the country and islands, it can be seen that bad farming and erosion have reduced the areas once cultivated. Wheat grew well enough, but, even as early as the 8th century BC, not in sufficient quantity to satisfy the demands of growing urban populations. Specialized crops producing oil or wine, which could be traded for wheat, proved a more economic use of land for some states (Athens and Chios, for example). The hillslopes provided pasture for cattle, sheep, and goats, which played an important part in the provision of food and clothing. Greece was well wooded, but the demands of housebuilding and shipbuilding, as well as the attentions of the goats, considerably diminished the forested area even in antiquity. There is good fishing in the Aegean, especially of tunny, and the Greeks ate much fish, fresh and dried.

Climatic changes since antiquity have probably been slight, but there may have been a period of drought in the Early Iron Age which contributed to the severe depopulation of the country in about the 12th and 11th centuries BC. There has been a general rise in sea level since antiquity of at least six to seven feet. Some parts of the coastline, notably in Crete and parts of the Peloponnese, have suffered local disturbance through earthquake. Most of Greece, especially southern Greece, is subject to these disturbances, which have at times proved disastrous. Mineral resources exploited in antiquity were few—some gold and silver and slight deposits of copper and iron. But there were rich sources of fine-quality stone for building and statuary and plentiful deposits of sedimentary clays for potting.

THE EARLY ARCHAIC AND ARCHAIC PERIODS

The story of ancient Greek civilization begins with the Early Archaic period; that is, the period from the disintegration (*c.* 1200 BC) of the Bronze Age civilization in Greece to the beginning of the literary and cultural revival of the Archaic period proper (*c.* 750–*c.* 500 BC).

The Early Archaic period. *The Dorian Invasion and Proto-Geometric period, c. 1100–900 BC.* The collapse and gradual destruction of the Bronze Age civilizations in Greece are subjects for archaeological investigation only, and it is natural to look first to archaeology for any account of the period that immediately followed. During this following period, the few remaining centres of the earlier Mycenaean culture were abandoned or destroyed. These included not only settlements at the older capitals, like Mycenae, but shorter lived refugee settlements in west Greece (Achaea, Cephallenia), east Attica (Perati); settlements on eastern Aegean islands (Emporio on Chios); and settlements in the islands that had hitherto escaped the worst effects of the invasions.

The "invasions" may be thought of as having occurred in waves for over a century. It is possible that many of the effects attributed to them were in fact the result of internal dissensions. Observing the areas that seem to have been least disturbed by the earliest attacks, and the directions in which refugees moved, it may be judged that the main course of invasion passed through Greece from the northwest. The identity of the earlier invaders is not abundantly evident in the archaeological record, which probably means that the attacks were carried out by raiders who may have quickly passed on or turned back with their prizes. Some objects of new types, or of types only familiar on the borders of the Mycenaean world, are introduced: an efficient bronze sword, violin-bow fibulae (clasps), some distinctive armour, knives, and handmade pottery. Prototypes for these are readily found in northwest Greece and the Balkans beyond.

The overall effect of the invasions was a depression of all cultural standards and considerable depopulation, except in Athens, parts of Thessaly and the Argolid, and in the refugee areas. It has been suggested that a contributory or even dominant factor was a climatic change that brought about conditions of drought and famine. The evidence for or against this hypothesis is difficult to judge, but the sudden upheaval, after so many years of prosperity and heavy settlement, is hard to attribute solely to the hand of man.

It is concerning the final wave of invasions or destructions in the early and middle 11th century that the archaeological record is somewhat more informative. It reveals some continuity of settlement with a new population or, at least, a somewhat changed way of life. This is most easily observed in Attica and the Argolid.

The new culture is distinguished from what had gone before in a number of ways. The Mycenaeans buried their dead in family graves. The new cemeteries are of individual burials in slab-lined cists. These are not altogether unknown in earlier Greece; they were normal in the Middle Bronze Age, and, toward the end of the Bronze Age,



Ancient Greece.

From W. Shepherd, *Historical Atlas*; Barnes and Noble, Inc.

cist cemeteries were made in north Greece (Epirus and Thessaly). As an alternative burial practice, cremation was introduced and became dominant in the following period in Attica and Crete but remained rare in the Peloponnese and is only occasionally found elsewhere. Cremation was rare in Mycenaean Greece but was practiced during the Mycenaean period in the central Balkans and was regular farther north. Straight dress pins make their appearance, and iron for weapons. These were to be characteristic of the new Proto-Geometric period, whose other characteristics are examined below. It is impossible not to associate the invasions of this period and the resettlement or continued settlement of the surviving Mycenaean sites with later Greek tradition about the Dorians. The Dorians were Greeks from northwest and north central Greece, and there is nothing in the archaeological record to upset the theory that they were the last invaders and settlers of the Mycenaean Greek world. They effected a shift of emphasis rather than a revolution in the Greek way of life. Tradition told of the "Return of the Heraclidae," suggesting earlier or repeated raids from the north. There were moves from Thessaly into central Greece and Boeotia and from Epirus and the northwest into the Peloponnese. Not all the immigrants were Dorian, but it was the Dorian invasion of the Peloponnese that overthrew the remaining Mycenaean centres and led to the strongest new settlements in Laconia, the Argolid, and Corinthia. Athens was marginally affected, perhaps mainly by refugee populations. A Mycenaean "pocket" remained in Arcadia, using a dialect related to that of the Mycenaean refugees in Cyprus. Dorians settled also in the southern Aegean islands, in Rhodes, and especially in Crete, where again there were pockets of Bronze Age survival—some even non-Greek Mycenaean (the Eteocretans). It is possible that less of Greece was totally depopulated during the invasions than has been thought. It may well be that the interaction between the new and old Greeks in what had been centres of Mycenaean power proved responsible for developments in religion and myth about which it is now only possible to speculate, since no contemporary written records survive.

Identification of the invaders

In Greece itself, the art of writing had been lost with the palace bureaucracies that disappeared as a consequence of the invasion. One of the prime achievements of the new Proto-Geometric period of which there is material evidence was in pottery and pottery decoration—from this the period gets its name. The notable refinement in shapes and decoration exercised by the Proto-Geometric potters represents not a break with earlier tradition but a reworking of old forms and patterns with a new precision and eye for proportion and design. This is the beginning of Greek Classical art. It seems most likely that the inception of the new style took place in Athens, and it was certainly in Athens that it was most successfully developed. And from Athens other areas of Greece drew inspiration. Athens was never Dorian, but it shared with the Argolid the first manifestations of the new style. Skeletal material from Athenian graves reveals a marked intrusion of northern types, and so the responsibility of the new Greeks for the changes observed in Athens may well be real, and possibly not wholly indirect.

Cemeteries are the main source of information for the Proto-Geometric period, and it can be seen from them

that as the style develops, through the 10th century, there is a growth of population accompanied by growth of prosperity. This prosperity is indicated by the increasing number of metal objects deposited in the graves. Apart from east central Greece and Argolis there was a strong unity of development in this period in Thessaly, Euboea, and into the Cyclades. In these areas the Proto-Geometric style died hard. The Euboeans and islanders, in particular, seem already at this time to have been busy seafarers; and, from the site of Xeropolis (Levkandí) on Euboea, between Eretria and Chalcis, there is evidence for a variety of imported goods from Cyprus and the Near East. By comparison, Laconia and western Greece were backward, slow to repopulate and to trade. Crete is a special case since the island suffered less than most of Greece from the disasters attending the end of the Bronze Age. Some centres in the island seem to have developed with a certain vigour and degree of independence. There are clear signs in Crete, also, of contact with Cyprus and the Near East.

Apart from archaeological sites, pottery, a restricted range of metalwork, and cemeteries, little can be assessed of Proto-Geometric culture. Virtually nothing is known about temples or even houses. Yet it was during this period that the tenets of myth-history and religion found in an already developed form in the Homeric poems at the end of the following period were being established. Much of the ordinary way of life described in those poems was probably typical of the Proto-Geometric period. And, despite an overall unity of culture in Greek-speaking lands, it is possible to detect that differentiation of customs (mainly burial) and style that came to be most fully expressed in the stout independence of the city-states.

The Ionian migration

There is one other event, rather loosely described and dated by ancient writers, which archaeology suggests may be placed in this period—the "migrations" to the east coast of the Aegean. The area had already been much visited and partly settled in the Bronze Age, and offshore islands had received refugees from the Mycenaean world. But none of these communities survived into the Proto-Geometric period, and for most of the area there is reason to believe that there was a break in Greek occupation, although not so obliterating that later Greeks were unable to find their way again to the same sites. The migration was described by ancient sources as a single event involving the marshalling in Attica of refugees from Pylos, together with other Greeks from Boeotia and elsewhere, and their passage east to found Miletus and Ephesus, and thence the other major cities of Ionia. There may be archaeological evidence for such refugees in Late Helladic Attica, and there is certainly archaeological evidence for the settlements in Ionia in the 11th and 10th centuries. The style of pottery found in them indicates some relationship with Attica. Miletus was founded by the start of the Proto-Geometric period, and this seems to support the literary tradition. Finds in most of the other major cities indicate that all were probably established before the end of the 10th century, but the diversity of the records about their composition and founders and the inadequate evidence of excavation make it difficult to decide how much was effected by fresh immigration from the homeland and how much by local expansion.

By about 900 BC Athens had emerged as the cultural centre of the Greek world; the Euboeans and islanders had begun their eastern voyages, which were to lead to new trading ventures and transform the prosperity of the Greek world; and the Aegean had become for the first time a truly Greek sea.

The world of Homer: the Geometric period, c. 900–c. 700 BC. The Geometric period is named after the style of pottery decoration that was developed in Athens out of the more austere Proto-Geometric idiom. It was wholly abstract at first but eventually admitted figure decoration. The closing phases of the period saw the introduction of elements derived from the arts of the Near East, a phenomenon that resulted from new Greek trading enterprise in the east Mediterranean. The desire and need for this trade reflect the more stable and prosperous conditions of life in Greece, which gave rise to a notable increase in population and the need for colonizing ventures. During the same period occurred the introduction of writing and the possibility of written records. Greece moves into the light of history.

The material culture of the new period is best judged in Athens, where the cemeteries encircling the city have proved most productive, and in the countryside of Attica. The method of burial becomes more varied, ordinary inhumation becoming as common as the cremation that had been dominant in the preceding period. Grave offerings, with weapons and metalwork, give some indication of the new standards of life, while some burials in Athens (the so-called Dipylon graves) are marked by monumental clay vases decorated with scenes of the laying out of the body and its removal to burial, implying funeral rites and celebrations of some complexity and expense. Iron appears more often and is used for swords and spearheads; bronze is used for dress pins, ornaments, and implements. There is a considerable increase in the production of minor works in bronze and clay for dedication in sanctuaries, mainly in the form of men and animals. The most impressive bronzes are the great tripod caldrons, best known at Olympia but also in Athens and elsewhere in Greece. Precious metal was still comparatively rare, but immigrant eastern goldsmiths were at work in Attica in the early 8th century, and there was some local production of goldwork in the Geometric style. Some eastern metalwork and jewelry had arrived by the 9th century and had been deposited in graves. Another luxury craft, forgotten since the Bronze Age, was the cutting of seals in ivory and stone, indicating some sophistication in the safeguarding of property and identity of ownership.

This Geometric culture of Athens is repeated all over Greece, although nowhere else with comparable complexity and wealth, except possibly in Crete. Athens still seemed to give the lead in pottery decoration, but regional schools developed, especially in the 8th century. The new wealth and growing population of the Greek states were soon to lead them to test their strength and independence against each other in fields other than that of craftsmanship.

The Proto-Geometric unit formed by Thessaly, Euboea, and the Cyclades remained a strong entity, with Euboea the dominant partner. At Levkandí, the site between Eretria and Chalcis in the Lelantine Plain, there is clear evidence for metalworking and signs of trade with the Near East. The town was destroyed before 700 BC, while the site of classical Eretria was first settled only in the early 8th century. These events are in some way to be related to the Lelantine War, a conflict between Eretria and Chalcis, which, according to ancient authors, split Greece into two camps. The Thessalians were, not unexpectedly, directly involved. The affiliation of Levkandí and the outcome of the conflict, which may have been protracted, are not clear, but the immediate result was considerable restriction in the activity of the Euboean states at home and overseas. In the islands there were several important Geometric towns, notably on the sacred island of Delos.

The Lelantine War

In Boeotia, Thebes seems the strongest site and most active culturally, although dependent on Athenian styles. Delphi was prosperous by the 8th century and received many Corinthian goods.

In the Peloponnese, archaeological evidence indicates the rapid growth in wealth and size of Corinth and a considerable concentration of prosperous sites in Argolis—at Argos, Mycenae, Tiryns, and Nauplia especially. While there is not the volume of imports and precious metals found in Attica, it is in Argos that is found the earliest suit of all-metal armour of the type later worn by hoplite soldiers, as well as offerings of spits (*obeloi*) in sets of six. Corinth's prosperity began in the 8th century, especially after the middle years, when its pottery began to appear in quantity on a great many Greek sites at home and overseas.

Crete presents a checkered picture, with several important sites in the centre (notably Knossos) and the east. Relations with Cyprus remained strong, and by the end of the 9th century a group of eastern goldsmiths had settled in Knossos. They worked there, in gold, bronze, and stone, for some three generations. Crete, like Thera—another Dorian island that prospered at this time—generally preferred cremation for its dead. Some tombs, though

used for the new rite, retain the form of the old Bronze Age *tholoi:* the Minoan heritage is imprinted in a number of ways on Archaic Cretan culture.

In east Greece (Asia Minor and the eastern Aegean islands), several towns on Dorian Rhodes were flourishing and developing a regular trade with Cyprus by the 9th century. It is not until the 8th century, however, that the major sites in Ionia show marked signs of growth, notably at Smyrna and on Samos, where the first temples of the Heraeum were built. The archaeological record from other cities and islands is sparse. Farther north, the coastline of Aeolis as far as Troy was settled. Finds on sites as far south as Chios and Samos show evidence for increasing awareness of the hinterland powers of Phrygia and Lydia, while farther south Greek fashions were penetrating Caria.

The material pattern of life in Geometric Greece is characterized by growth in population and wealth, but a special quality is also increasingly apparent, and this must be summarized and explained. Objects from Cyprus, Syria, and Phoenicia (and to a limited degree Egypt) arrived in the Greek world in increasing numbers through these two centuries. The presence of immigrant craftsmen can be detected in Attica and Crete; there may have been ivory workshops in the Peloponnese, Attica, and east Greece; and, by the second half of the 8th century, most Greek crafts were beginning to betray the imminence of the influence from the East that became more fully apparent in the 7th century. This appears most clearly in Corinth, Crete, Rhodes, and Attica. The goods and craftsmen followed routes that may not at first have carried much else by way of serious trade, but the expanding Greek states looked overseas for new resources, especially for metals—copper, tin, iron—that were in short supply at home and that were especially required for weapons. Tradition has it that the Phoenicians conducted much of the early trade with Greece, but physical evidence for this is hard to find, and the first clear sign of organized trade is a Greek enterprise. This involved the establishment of a trading post at the mouth of the River Orontes in north Syria at a site known as al-Mīnā. There is no literary record of the venture, but it, or a similar operation, provided Greece with its first opportunity to keep written records, since it was through intercourse with the Aramaic-writing peoples on this coast that the Greeks learned and adapted for their own use alphabetic writing. This must have happened by the mid-8th century, and if Olympia had accurate written records going back to 776 BC, then it may be surmised that writing reached the Peloponnese about that time.

Eastern influences on the Greeks

The physical world of the Homeric poems is dominantly that of Geometric Greece, overlaid as it is with the myth-history and geography of an earlier age and with stock epithets and descriptions perhaps out-of-date by the 8th century. The picture is a convincing one of independent kingdoms centred on single cities with dependent villages and linked only by bonds of kinship or mutual convenience. Rule is generally autocratic and fighting an affair of the nobility, the citizen mass not yet being a decisive force. Hesiod, writing around 700 BC, paints a more sombre but moralizing picture of the hardships of a farmer's life in Boeotia.

Homer and Hesiod are poor guides to religious life in Geometric Greece. Homer's gods are already acquiring set attributes and an Olympian family life, imposed in the interest of the story. The process may have been going on for some time, but this is bards' work and for long afterwards bore little relationship to the Greeks' view of the town's patron deity and that deity's functions and dues. Hesiod's cosmogony is, if anything, even more artificial and owes much to Eastern cosmic myths. His systems and tales also find little echo in later Greek religious practice, but in parts of his work natural justice and the will of the Father of the Gods correspond well enough to permit recognition of an attitude to the gods based on humane respect rather than fear. Archaeology can offer evidence of only the trappings of religious life. The earliest cult places are altars for sacrifice, which also serve as the focus for offerings. The provision of a house (*oikos*) to accommodate the symbol or cult image of the deity comes late, and the few Geometric examples of such primitive temples all resemble contemporary domestic architecture. The images themselves may well have been aniconic (symbolic rather than representational). The offerings are bronze or clay substitutes for the offerer himself, representations of the deity, or intrinsically valuable objects such as the metal caldrons that in the Homeric poems are offered as prizes. In the grave a warrior may be buried with his weapons, sometimes rendered useless (swords bent or folded), but most offerings are of pottery and seem to indicate provision for a meal, no doubt for the journey to the other world. There is no clear evidence for continued rites at the graves of the recent dead, but in the 8th century there is evidence of attention being paid to some Mycenaean tombs and of the introduction of some hero cults which may have been inspired by the Homeric poems, as at the Agamemnoneion of Mycenae.

Religion in the Geometric period

Certain other aspects of Greece's remarkable Bronze Age past must have affected its new and rapidly developing Iron Age civilization. The walls of Mycenaean citadels were viewed as the work of giants, the Cyclopes. These physical monuments, together with the poems, were clear evidence for a golden age, long past and replaced by one of iron. Geometric and later artists were very rarely moved to imitate these works. The degree of continuity from the Bronze Age is more difficult to assess but has been enthusiastically championed. The palatial arts were lost and had to be relearned from the East. Old sites reoccupied had led scholars to believe in the continuous use of cult places and sanctuaries, but they are difficult to identify in Bronze Age terms.

Although their new literacy may have awakened an intelligent interest in their past, the strongest motivation for progress in the Geometric period came from the Greeks' own efforts to improve their lot at home and from their enterprise overseas, which brought them both valuable new resources and invigorating new ideas.

The Greek colonizing movement, c. 750–c. 500 BC. *The Mediterranean world and Greek needs.* The geography of Greece determined the Greeks' interest in seafaring, and, when their homeland could no longer support their growing population, it was their seafarers who offered the solution. Early sailing was a hazardous business, generally restricted to a few months in the year and to routes which, so far as possible, never took the mariner too far from sight of land. Not until the 6th century were maps being drawn, and, even then, they were of little practical value, compared to experience and the reports of others. The need now was for land and new resources. What could the Mediterranean seaboard offer? First, a climate that was at no point so markedly unlike that of Greece that Greeks could not readily adapt themselves to the new conditions of life and agriculture, and, second, sea and river routes to remote peoples and resources. The sources of raw materials were discovered early and exploited fairly easily by trade, even if in the teeth of foreign, usually Phoenician, competition. The western Mediterranean gave access to the tin and silver of Spain, France, and even Britain; central and north Italy could offer copper and the iron of Elba. The richer mineral resources of Europe were reached only with more difficulty. On the shores of the Black Sea there was gold (in Colchis), and there was iron in north Asia Minor. The copper island, Cyprus, pointed the way to areas in which metallurgy had long flourished and from which many even more distant sources had been exploited. But food was even more important than metal, and this involved seeking new land for settlers. This, of course, depended upon finding unoccupied but cultivable land or upon the goodwill of new neighbours.

Expansion and trade

The Mediterranean world offered several such opportunities for expansion and trade. The northern shore of the Aegean flanked the territories of the minor kingdoms within Macedonia and Thrace. The Macedonians were Greek in culture; but neither they nor the Thracians were seafarers, and the coast of Macedonia, especially the peninsula of Chalcidice, proved an important colonial area for the earliest of the active colonizing cities in Euboea. Farther east, new towns were founded from the islands and east Greece, and the gold of Mt. Pangaeum was soon exploited. On the east shores of the Aegean the hinterland

powers also had little interest in the coast, and the already established Aeolic and Ionian cities soon expanded along the coastal strip up to the Dardanelles. The passage up into the Black Sea, against wind and current, was difficult but not impossible. The Thracians to the west seem to have welcomed the Greeks. In the north the newly arrived nomadic Royal Scythians provided a ready market rather than competition, and again the Greeks were able to benefit from the fact that they were more interested in the sea and coastal areas than the natives were. They soon learned to exploit the riches of the south Russian grain fields and fisheries. The less hospitable eastern and southern shores of the Black Sea offered fewer but important places for foundations that could tap important mineral resources.

In Cyprus, the Near East, and Egypt the situation was very different. All the local states were well established, and all had an interest in seagoing. The Egyptians also had a deep-seated suspicion of all foreigners. Their trading routes inland were well trodden, and there was no foothold there for a colony. Libya was a different matter, with a broad green plateau and disorganized native tribes that appeared cooperative. Farther west, the Phoenicians had forestalled the Greeks, and Carthage was founded while the Greeks were still exploring Italy and Sicily.

The Greeks were slow to explore routes up the Adriatic and crossed straight to south Italy and up to the promising markets of Etruria. There, as in Egypt, the opportunities were for trade. The natives of south Italy and Sicily were landlubbers, their natural ports and coastal plains mainly empty, it seems. There the Euboeans, followed by Corinth and other Peloponnesian states, were able to establish important agricultural communities. East Greeks later prospected farther afield, to France and Spain.

There were solutions to the land problem other than the establishment of colonies, but resorting to these was rare. Sparta annexed the land of its neighbour, Messenia, acquiring a large slave population and its attendant problems. A dissident group in Sparta threatened state unity and was dispatched to found Tarentum in south Italy. Athens had been an active seagoer earlier in the 8th century, but it did not join the first colonizing movement and later avoided the necessity by specializing its economy in export materials—oil, wine, and silver. The Euboean cities, Eretria and Chalcis, were the earliest and most energetic traders and also the earliest colonizers, since their home island offered little scope for their new wealth.

Relations between colony and mother city

The first colonizing parties were probably very small, and families from neighbouring states seem often to have been free to join. This contributed to mixed feelings of loyalty, and, given the forced circumstances under which some colonies were dispatched, it is not surprising that relations with mother cities were sometimes less than cordial. When the colonizing states were seagoers, like the Euboeans and Milesians, the routes and possible destinations were already known to their merchants. The Therans turned to a Cretan who knew it was safe to bear south for the Libyan coast. Other states may have sought information from the Delphic oracle: the latter's newfound wealth and reputation may have been in some part due to the role it played or claimed as promoter of the siting, cults, and laws of new foundations.

The immediate needs of any new city may have been quite basic, but its identity as a full Greek *polis*, or city-state, on the model of its mother city, was immediately affirmed by the establishment of the appropriate cult places and the demarcation of an agora area for political and commercial assembly. The new cities had not the architecture of an old city to impede or limit their planning, and their political and legal institutions were perhaps less trammelled by local, family, or territorial traditions. But it did not take long, with the growth of wealth and strength, for their political and social life to acquire all the complexity of the homeland.

The trading posts. There had been refugee Mycenaean Greek settlements in the island of Cyprus, and some surviving Greek element in the culture and population of the island may account for the fact that it was only with Cyprus that some sort of trade relations, even if casual, were maintained through the "dark ages" of Proto-Geometric and Geometric Greece. Archaeology shows that this was true especially of Crete, Euboea, and eventually Athens. Cyprus itself, from its geographical position, was a cultural extension of the Levant, and, by the 10th century at least, new Phoenician cities had been founded in the island. Before the end of the 9th century it seems that Cypriots had led Euboean traders to the eastern mainland and, with them, established a trading port at the mouth of the River Orontes, at a site known as al-Mīnā, excavated before World War II but not recorded by any ancient author. Euboean and island interest was dominant there until about 700 BC, and this site and port, with immediate access to the Syrian hinterland and Mesopotamia, may have been the major source for the flow of Eastern materials into the Greek world of the 8th century. After 700 the Cypriot and Euboean interest gave place to one that was mainly east Greek but shared by a mainland Greek state, perhaps Corinth or Aegina—at least a state using Corinthian pottery. By this time there may have been Greek trader elements in other eastern cities, such as Tarsus in Cilicia and farther south on the Syro-Phoenician coast (at Tall Sūkās, for example). There would have been no serious rival to this route, which brought prosperity to another Greek staging point for Eastern objects destined for the Greek world—Rhodes. In Asia Minor, the roles of Phrygia and Lydia as purveyors of Eastern goods and ideas seem to have been strictly limited to the east Greek world, and the direction of the flow was soon turned.

In Cyprus and north Syria there was no question of establishing colonies, although in time Cyprus was to become thoroughly Hellenized and isolated from its Eastern associations. Greek trading in Egypt gave rise to a still different type of settlement. In the mid-7th century the Egyptian king Psamtik I had employed east Greek mercenary soldiers, some of whom were allowed to settle in the country, and it is evident from excavations that by about 620 BC a Greek trading town had been established at Naukratis on the western branch of the Nile Delta. Herodotus describes a reorganization of the town under King Ahmose II in the mid-6th century that confirms what the finds indicate: that it was in the hands of various east Greek states and of Aegina. This was virtually the same consortium that had been operating at al-Mīnā since about 700 BC. In Egypt their operations were limited to trade through Naukratis, carrying in wine, oil, and silver, and carrying off papyrus, manufactured goods, and, eventually, grain. The community was better established than at al-Mīnā, since the Egyptians allowed a degree of local government by overseers appointed by each trading state and a number of temples, including a joint Hellenium. The latter may have been involved in collecting the taxes that had to be paid over to Egyptian temple treasuries.

The Greek trading settlement in Egypt

Precolonial trade in other areas of the Mediterranean world was far less firmly established and was conducted, it seems, by regular visit rather than through trade settlements. Evidence for it, therefore, is in the form of Greek objects in native towns. These appear sporadically in south Russia, up the Dnepr and the Don, and more often in Etruria. Some Greek place-names in the Carthage area suggest that Euboeans may have explored this coast, too, but Phoenician settlement there by about 750 BC forestalled possible development of trade or settlement.

Greek colonies in the west. Finds in Etruria show that Euboean merchants had been visiting its shores early in the 8th century BC. Before the middle of the century, a settlement had been established on the island of Ischia, offshore of Naples, at a site called Pithecussae. It was easily defensible and offered a good harbour at a convenient point on the approach to the Etruscan cities in the north. Its early years were its most prosperous, it seems, and the Greeks on this agriculturally unpromising volcanic island must have relied heavily on trade. There are early traces of ironworking, with material no doubt acquired from Elba. By the middle of the century, a second colony was established, on the mainland at Cumae, mainly by Chalcidians from Euboea. This provided better opportunities for livelihood off the fertile coastal strip, and it naturally came to eclipse Pithecussae in importance. These, the first of the Greek colonies, were also the most remote on this

coast and the nearest to Etruria. The approach to them from Greece lay through the Strait of Messina, and the Chalcidians soon took note of the Sicilian coast—both for staging points and for its resources. The Chalcidians founded Naxos in Sicily in 734 BC. It seems that the Greeks had to displace a native settlement, and the same was true of the Chalcidians' next foundation in Sicily, at Leontini, which lay to the south of Naxos and a little inland, commanding a rich plain. Lower down the valley, on the sea, they founded Catana in 729 BC. This had a good harbour and plain. Thus it can be seen that, as in Italy, the first settlements were determined as much as anything by marine or mercantile needs and were soon supported by agricultural foundations.

The passage through the Strait of Messina had also to be safeguarded. Chalcidians, with a party from Cumae, founded Zancle (later called Messana, modern Messina) on the Sicilian side, in a position that was more strategic than comfortable and that was supported in 716 BC by the foundation of Mylae, at the other side of the northeast spur of Sicily, where there was a plain. On the Italian side, Rhegium was founded at about the same date by Chalcidians and a party invited from Messenia in the Peloponnese (noble refugees from Spartan expansion). The Euboeans had sought to safeguard their approach to the west by a colony on the western Greek island of Corcyra. In 733 BC the Eretrians were expelled from the island by Corinth, whose own settlers were soon at loggerheads with their mother city. They did, however, eventually press on up the Adriatic, founding Epidamnus in Illyria in 627 BC. Corinth's eyes were not on the north, however, but on the west and the trail blazed by the Euboeans, and its further colonizing along the western coast of Greece was mainly in the interest of protecting the approaches to the Corinthian Gulf.

The founding of Syracuse

According to ancient sources, it was only a year after the first Euboean foundation in Sicily that Corinthian settlers occupied the best port site on the east coast, and one which was to become one of the strongest cities in the whole Greek world—Syracuse (*c.* 733 BC). This was, in fact, an island site, soon linked to the mainland by a causeway; it commanded a broad coastal plain that seems to have been virtually unoccupied. Settlers from Megara had a harder time, spending a while in Chalcidian Leontini, then moving to a bare promontory at Thapsus before being invited by the native Siculi (Sicels) to what was to prove a valuable coastal site—Megara Hyblaea—in 728 BC. All these foundations were in the east of the island, and it was another generation before some Dorians, from Rhodes and Crete, ventured a little way along the south coast, to Gela, in 688 BC. Farther west lay the Elymi (Elymians), but active Phoenician interest in the west of the island and the establishment of a Phoenician town at Motya some time before 700 BC discouraged the Greeks from pressing far in this direction, although there was at this time no direct confrontation between the two great colonizing and trading nations.

The Euboeans established colonies in Italy that were as much strategic as agricultural. They were soon followed by Peloponnesians who had an eye rather to the rich and unoccupied coastal plains that the Euboeans sailed past. Most came from Achaean cities. Sybaris, whose prosperity was short-lived but became legendary, was founded in about 720 BC, followed by colonies at Metapontum, Croton, and eventually Caulonia. These enjoyed the best of the land on the southeast coast, but there were valley routes northwest to the Tyrrhenian Sea, bypassing the Strait of Messina, and there was an early flow of Greek goods through the native towns of the hinterland. The other important Achaean colony, Poseidonia (Paestum), lay at the end of one of these routes and was established by about 700 BC.

The Achaeans were not alone in south Italy. In about 706 BC the Spartans settled the fine port at Tarentum, beside a native town that had been in touch with Greeks long before. Locri was settled from central Greece in about 673 BC. The first Ionian colony was at Siris, (later called Heraclea) settled by Colophonians who had been displaced from their homes by Lydians—a circumstance that was to set other Ionians on the move west in the next century. Within less than a century, and mainly in its first 50 years, Greek states had found for themselves new and prosperous homes on the coasts of south Italy and Sicily. They had established generally good relations with their neighbours, were well placed to dominate trade with Etruria, and were already courting the jealousy of the Phoenicians and Carthage.

The first colonies had all been on or very close to the sea. The period of expansion and consolidation, after the mid-7th century, saw more inland foundations, bringing the Greeks into closer contact with the native peoples. New cities toward western Sicily began to challenge the Phoenician hold on that part of the island. Although some of the new cities were manned by newcomers from Greece, much was undertaken by the colonies themselves, which sought to confirm and safeguard their positions.

Expansion of Syracuse

Syracuse established its hold on southeastern Sicily by founding Acrae (663 BC); Casmenae, yet farther inland (643 BC); and Camarina (598 BC). Smaller settlements, as at Helorum on the coast south of Syracuse, confirmed its domination of the grain lands and, by enslaving the Siculi, Syracuse had, by the early 6th century, progressed from being an outpost on a foreign shore to command of an area and cities rivalling any controlled by a homeland city. Camarina alone claimed a measure of independence: it quarrelled with Syracuse and was defeated and laid waste in 552/551 BC.

Megara Hyblaea, hemmed in by the Chalcidian cities and Syracuse, looked to the opposite end of the island and founded Selinus soon after the mid-7th century in territory close to the Phoenician town of Motya. This was to prove another of the most prosperous western cities, long to outlive its mother city. At about the same time (648 BC) Chalcidians, joined by a group from Syracuse, also struck westward to found Himera, lying about as close to Phoenician Panormus (Palermo) as Selinus did to Motya.

The last of the important Sicilian foundations was made by Gela at Acragas (Roman Agrigentum) along the coast to the west, in 580 BC. Eastward expansion was contained by Syracuse, and the Geloan cities were to be a serious challenge to Syracusan supremacy in Sicily.

In south Italy, the new foundations were mainly of the 6th century and were sponsored by the existing colonies establishing routes across the peninsula or "filling in" the coastline. The only newcomers were from east Greece. They had considerable trouble establishing themselves and provoked the first serious troubles with the Phoenicians. A party from Rhodes and Cnidus tried to settle too close to the Phoenicians at Lilybaeum in west Sicily and were expelled soon after 580 BC following a dispute involving the Phoenicians, Selinus, and Segesta—a native Elymian city that was strongly Hellenized and constantly at odds with Selinus. The unlucky colonists moved on to the Lipari Islands, where they established some form of communal ownership of land and annoyed the Etruscans by attacks on their shipping.

In the face of the Persian invasion of western Asia Minor in the mid-6th century, several east Greek states considered the possibility of mass emigration to the west, one suggestion being that the Ionians should move to Sardinia. The Phocaeans lost their city to Persia and sailed west to join the colony they had founded at Alalia on Corsica in 560 BC. Their piratic behaviour produced an alliance against them between the Etruscans and the Carthaginians, and, although the Phocaeans defeated them at sea, they had to leave Alalia and seek a new home. They stayed a while at Rhegium and then founded Elea (Velia) on the coast just south of Poseidonia in about 535 BC.

Political life of the colonies

Internally, the Greek colonies developed much as their mother cities at home. There were rare examples of political innovations, as at Lipara. The progress was generally from aristocracy to rule by tyrants, but the landowners of Syracuse were long influential. Disputes between the colonies did not always follow "national" lines, and the division between Dorian and Ionian had little weight in the colonial period.

Relationships with the native population rapidly became a matter of master and slave in Sicily, where only the

Elymians enjoyed some security under the protection of the Phoenicians in the west. The other native towns of Sicily became progressively more Hellenized. Many were taken over, so that by the end of the 6th century the large native population was no longer a significant force in the island. Much the same was true in the southernmost part of the mainland, but the native tribes of Apulia, Lucania, and Campania were made of stouter stuff, retaining an individuality of culture well into the Classical period.

Conservatism is perhaps a characteristic feature of colonial areas. It was certainly a feature of the culture of the western Greeks. Their new cities were rich and powerful, and they sought to demonstrate this wealth both by the splendour of their offerings and by athletic missions to the national sanctuaries and games of Greece itself, and by their building at home. At Olympia at least half the pavilion treasuries dedicated by individual states were paid for by westerners—Gela, Metapontum, Sybaris, Selinus, and Syracuse. A majority of the surviving temples of the Greek world constructed in this period stand in the west, and in Syracuse a start was made to emulate even the colossal temples of the Ionian cities. This was an isolated venture in the Ionic order: most western work is Doric, massive and conservative in proportions, but often admitting uncanonical decorative detail, "vulgar" by metropolitan standards. Lacking fine white marble, the statuary is in stuccoed limestone or clay and is decidedly provincial in appearance. Only in the minor arts of clay and bronze did some western studios make original contributions to the development of Greek art.

The Phoenicians had been trading with the south of Spain since the later 8th century, and by 700 BC they had established at least one trading post there. The Greeks also took an early interest in these western waters and for a while held their own against the Phoenicians, whose foundations in Africa, Spain, Sardinia, and western Sicily eventually led to their domination of this area. The attraction was the tin and silver obtainable in Spain and, via Spain, from more northerly sources. Greek goods continued to arrive in Spain through the 6th century, but many, like the earliest of them found in the Phoenician trading post at Almunecar, may have been carried by Phoenician ships.

An alternative route for the tin was overland, through France, down the valley of the Rhône to Marseille. Here the Phocaeans, who seem to have been the most active of the east Greek merchants in these seas, founded Massilia in about 600 BC. A vigorous trade into France and north Spain developed and was later reinforced by other colonial foundations from Massilia. By this route Greek works of art travelled deep into France and Germany, and a Celtic chieftain at Heuneburg near Munich seems to have employed a Greek to plan his fortification for him. But, by the end of the 6th century, the balance of power had shifted, and, as a result of the diminished trade along the Rhône, Massilia's days of high prosperity were over. Its fortunes are an example of how much and how long some colonies relied on trade rather than on local resources for their riches and even livelihood.

Greek colonies in the north. The Euboeans, who were the first in trade and colonizing in east and west, were also the first to exploit the northern shores of the Aegean, where their neighbours were the part-Greek Macedonians. The centres of activity were the lower valleys of the Axius and Strymon and the three-pronged peninsula between them that took its name, Chalcidice, from one of the colonizing cities. It seems that both Eretria and Chalcis had been active in these regions in the 8th century: the Chalcidians at Torone and Dicaea; the Eretrians at Mende and then at Methone. After the Lelantine War Chalcidian interest was dominant in this area, and in the mid-7th century Chalcis sponsored colonies sent out from the island of Andros to sites on the east coast of Chalcidice—Acanthus, Argilus, Stagirus, and Sane. Eretrian control seems to have been confined to the westernmost promontory, Pallene. It is likely, however, that it was by agreement with Chalcis rather than Eretria that Corinth established Potidaea at the neck of the Pallene peninsula in about 600 BC.

Colonization of Chalcidice

The important Euboean foundations had probably been made by about 700 BC. In the 7th century, when the coast to the east was being colonized, it was other Greek states in the islands and east Greece that were active. An early foundation, *c.* 680 BC, was by Paros on the island of Thasos. Thasos soon set colonies on the mainland opposite, at Neapolis (Kaválla) and Oisyme. From Ionia the Chians came to colonize Maroneia and the Clazomenians to colonize Abdera, on the coast east of Thasos. There is indication of serious trouble with the natives—Thracians in this area, belligerent and reluctant to Hellenize. They expelled the Greeks from Abdera for a while, but, after the mid-6th century, it was settled again from Teos in Ionia, by Greeks escaping from the Persian threat. These settlements had a trade interest in the minerals and timber of the mainland, but they were also strong agricultural communities where the vine was grown with considerable success.

Dates as early as the 8th century BC, which have been proposed for some of the Black Sea colonies, are supported neither by finds nor by considerations of geography, which show Greek colonial interest in the approaches to the Black Sea only at the end of the 7th century. But references and names in Greek poems show knowledge of the area by about 700, as in the Argonaut story. Miletus was the leading colonizer in this area, and most of its foundations were made in the later years of the 7th century, sometimes with the collaboration of other Ionian cities. The progress was not one of gradual exploration along the shores but of settlement, within a single generation, of key sites around the whole of the inland sea, especially at river mouths and points propitious for trade with the natives. The greatest Milesian city was at Olbia, at the mouth of the River Bug, but settlement there seems to have been preceded by a post on the island of Berezan, farther out in the Bug-Dnepr estuary. Olbia was the main point of contact with the Scythians of the south Russian steppes. Herodotus tells of the Scythian prince who was so corrupted by the Greek way of life that he was murdered by his followers.

South of the Danube delta the Milesians founded Istrus, and later Tomis (modern Constanţa, Romania) to the south, on what was to be a frontier of the Roman Empire. The other important area of Milesian colonization lay around the Kerch Straits, leading into the Sea of Azov east of the Crimean Peninsula. The oldest and strongest city in this area was at Panticapaeum (Kerch).

On the east coast, the wealth of the Caucasus was known to Greeks, and there was settlement there by Milesians at Phasis and elsewhere, about which there is still much to learn. The coastal route back to the Aegean lay along the north shore of Asia Minor, which is notoriously inhospitable to shipping. Three important staging points were settled as colonies along this coast, at Sinope, Amisus, and Trapezus (Trebizond). Apart from their value to transit shipping, Amisus, in particular, also gave access to the mines of Anatolia.

The Megarians, from mainland Greece, were other notable colonizers in the Black Sea region. They placed Mesembria to the north, and Heraclea just east of the straits, and in the 5th century crossed the sea to Chersonesus in the Crimea.

Economy of the Black Sea colonies

The prosperity of the Black Sea colonies depended on various factors. They generally had no lack of land for local needs and were able to introduce Mediterranean plants without too great difficulty—the climate of the Crimea can be remarkably like that of the Aegean for much of the year. They relied, too, on their trade in the products or raw materials of the hinterland. Grain from the Black Sea was already on its way to Greece before the Greco-Persian Wars, and the local production of works of art in precious metals in some of the Greek cities is a fair indication of how much of the same material might have been sent home. The waters of the Black Sea also yielded their riches. Kerch herring were prized, and quantities of pickled fish were sent back to Greece. In the Scythians they had to deal with a people whose culture was in many respects barely inferior to their own. Although passage of goods was free between them, the Greeks never really succeeded in assimilating the Scythian, in either their arts or manner of life, in the way in which they could the

Eastern or the Egyptian. The frozen north remained for the Greeks the home of ogres and gifted savages.

Greek colonies in Africa. The easy route south from Crete to the Libyan coast may well have been travelled by Greeks in the Bronze Age. When, in the mid-7th century, the island of Thera was obliged to send away part of its population, it was from its Dorian neighbour Crete that a pilot was found to lead them south. They settled for two years on an offshore island and were visited there by Colaeus the Samian on his famous voyage west. They then moved for six years onto the mainland, to Aziris, before being led by the apparently friendly natives up on to the Libyan plateau where they found a site whose climate and setting closely resembled Greece, at Cyrene (about 632 BC). The attraction was agricultural, since the plateau was fertile and well watered. It was soon to become famous as a breeding ground for horses and as a nursery of the valuable silphium plant (now extinct), and in time it became an important source of grain for the Greek and Roman world. The foundation at Cyrene was soon followed by exploration along the coast. Apollonia was founded as the port of Cyrene, and to the west, the site later called Ptolemais, Taucheira, and Euesperides (modern Banghāzī). All of these, with the possible exception of the last, were founded before the end of the 7th century. The royal family at Cyrene was not always popular, and in the mid-6th century Barce was founded by an oligarchic faction, together with brothers of the king, farther west on the plateau.

This was the one colonial area that early suffered at the hands of the Persians. The cities submitted to Persian rule once Egypt had fallen to the new empire in 525 BC and had to suffer one armed Persian raid as far as Euesperides. But, even in this period, it is clear that Cyrene prospered, as its temples and tombs matched those of some of the richer western colonies. Dorian interest in the area was maintained not only by the founder islands, Thera, Crete, and Rhodes, but also from the Peloponnese, notably by Sparta.

In Egypt itself, the fortunes of the Greek trading town at Naukratis flourished through the 6th century and were barely dimmed by the arrival of the Persians. There were other small Greek communities in Egyptian towns that may have admitted "native quarters" or mercenary garrisons, but never anything like a colony. Out in the Libyan desert, at Siwa Oasis, the cult of Amon attracted Greek attention and soon became a recognized and respected oracle for all Greeks. The god was recognized as Zeus Ammon and figures on coins of Cyrene.

To the west, any Greek exploration had been cut short or wholly forestalled by the Phoenician settlement at Carthage. The only Greek attempt in this area came soon after 515 BC, when the Spartan Dorieus was led by Therans to make a settlement at Cinyps on the bare coast of Tripolitania. Here he was pressed by both Phoenicians and natives and had to retire.

Archaic Greek culture. *Religion in Archaic Greece.* The Archaic period offers a first opportunity to study more closely the religious life of the Greeks without paying undue attention to the rather artificial and literary pantheon of the Olympian family and dynasties of heroes. These, however, still provided the material for song and epic and were increasingly invoked or manipulated to serve political or personal ends. In the city-state can be distinguished two elements: the agora, which served as assembly and market place, and the temple area, where its god or gods were honoured. All the monumental architecture of Greece in this period was devoted to temples, altars, and other sanctuary buildings, and only toward the end, in Athens, were public administrative buildings treated in a comparable manner. The major deity of a city may have been determined by its Bronze Age history and the identification of a warrior god or fertility goddess as one of the Olympian pantheon; or by the promotion of a local deity or hero, often associated with a natural feature, to the same rank; or, in a colony, by the precept of the mother city or the Delphic oracle. The temple was at once a statement of confidence in the guardian power and a demonstration of wealth. Loyalty to the city god could be a stronger incentive to concerted action than loyalty to a family or tyrant. And, as the national sanctuaries won greater importance, either through their oracles or the staging of games, so loyalty to their deities—the Zeus of Olympia, the Apollo of Delphi—strengthened, together with the bond of language and their attitude to the "barbarian."

Dominance of religious architecture

Festivals were of different types and could serve different purposes. The state festivals, notably the Panathenaea in Athens, were occasions for the display of citizen solidarity and were readily exploited by tyrants to reinforce their own position. The many other festivals generally had a connection to agriculture—spring or harvest festivals, or celebrations associated with particular crops or activities (Dionysus for wine, Artemis for hunting)—or to the hero of a locality. The agricultural character of these is apparent in accounts and representations of the Archaic period, but their significance gradually became overlaid by other considerations when they were performed by city dwellers, whose life and interests became more and more remote from the countryside.

The conduct of festivals and of sacrifices by private individuals for favours received or hoped for was in the hands of priests whose office went by lot or family. Burial was a private affair but was regulated by the state. The Homeric magnificence of burials of the Geometric period was restricted by legislation (as by Solon in Athens) generally directed against excessive expense and public display. In the Archaic period, the dead were better served above ground, by monuments that reflected as much credit on the living, than below; and the few offerings laid with the body were but a token of a presumed brief need for sustenance on the way to the other world. But some religions or sects provided for man's fears for his afterlife. This was achieved through ritual purity in way of life or through initiation in a "mystery" religion such as that at Eleusis. The former method involved a complicated theory of evolution and matter and a mythology of the underworld with precepts for the journey there. Such was Orphism. The philosophy of Pythagoras, who left Samos to found a school in Croton in the later 6th century, was similar, teaching the migration of the soul into another body, which might be human or animal. Pythagoreanism codified many precepts that elsewhere would be classified as magic. Orphism, together with aspects of the Dionysiac worship (which was late to win importance in Greece), was also rooted in a religion of ecstatic "possession" by the god working through a priest or in the bodies of dancing worshippers. In the Archaic period there was, thus, little in the religious practices of the Greeks to mark them from other early societies. The difference lay in their literary treatment of stories about gods, from Homer on. Only with the Classical period did their interest in man and man's behaviour begin seriously to condition their views about their gods.

Orphism and Pythagoreanism

The early philosophers. The cosmogonies of Hesiod and other early poets described the beginning of the world in terms of the supernatural, but in poems that, at the same time, showed the artists' shrewd observation and appreciation of man as an object deserving study. The sometimes ribald treatment of divine behaviour probably accelerated the process of critical inquiry and scepticism about the validity of myth and the attempt to replace it with a view of the world and man's place in it that did not depend upon the antics of anthropomorphic deities. In the 6th century the centre of this new criticism lay in Ionia, and especially in Miletus, the city of explorers. The exercise of reason and the seeking of truth through argument were the two methods of inquiry, and they owed virtually nothing to Eastern science and very little to scientific observation. This is why the systems devised by the early philosophers seem so naïve and impractical to the modern mind, but what is most interesting is why they saw the need to devise them at all. They looked for the basic principles of matter and the world around them—in water (Thales); in the "boundless" separating into opposites (Anaximander); in air (Anaximenes); and fire (Heracleitus). Yet they were practical men, too, involved in the commerce and politics of their day, and it was probably due to their understanding of some of the mathematical and geometrical

discoveries of the East and Egypt that certain of the more ambitious technical achievements of late Archaic Greece were made possible.

The Archaic style in the arts. The Archaic period saw the absorption by Greek artists of the artistic styles and techniques of the Near East and Egypt and their molding into a distinctive and original idiom in which lay the seeds of everything that was to be achieved in the Classical period.

During the 8th century foreign craftsmen and goods had come to Greece, but only toward the end of the century did these new arrivals begin to exert a profound effect on the work of Greek artists. Their metallurgy was always of a high order, and they refined techniques of casting and hammering bronze long before this was practiced in the East. New materials required new techniques, as in the case of ivory, and this, together with the filigree and granulation of the goldsmith, was a craft that had to be learned afresh, although it had flourished in Bronze Age Greece. The formal changes were more important. Geometric art could be subtle, but the more realistic conventions of Eastern art, involving the rendering of detail, offered far greater possibilities in the depiction of action. This was not, however, immediately applied to the needs of narrative or myth, and the hallmark of Eastern-influenced Greek work is the appearance of animal-frieze decoration. With the animals came a selection of monster motifs like the sphinx and griffin, to which the Greeks could add their own subhuman creations like the centaur and the satyr. Another new element was floral decoration, which gradually replaced the geometric and offered the possibility of considerable compositional invention—of which the Greeks were not slow to take advantage.

Developments in sculpture

In the mid-7th century the Greek feeling for the monumental was stimulated by knowledge of Egyptian art. This gave rise to the first sculpture in hard stone (white marble) at life size or more, which successfully overcame the sterile conventions of work done in more easily managed materials and at a smaller, purely decorative scale. It also led to the creation of monumental stone architecture, with "orders" of moldings and columns based on timber prototypes (Doric) or on Oriental patterns (Ionic). This successful expression of the monumental probably helped the vase painter shake off the more restrictive conventions of decoration and indulge wholeheartedly in scenes of everyday life or myth, as on Athenian black-figure vases of the 6th century. A set iconography of scenes and figures was soon established, but never so rigidly as not to allow for the invention of a gifted artist or for originality in the execution of detail.

Progress in the physical appearance of figures in Greek art through this period was not rapid, despite the ease with which technical problems had been overcome. The figures were rendered with greater detail, and this was inevitably more realistic. Overriding considerations were proportion and pattern—the pattern of anatomical detail or the pattern of drapery pulled tight across limbs or hanging free from the body in a cascade of folds. In representing poses, the artist only very slowly broke with the uncompromising frontal or profile views and the abrupt juxtaposition of these views. But the change was not toward greater realism for its own sake. It was rather a refinement of the expression, by means of symbols for heroes, men, or monsters, of more and more complicated action and compositions. It was only just before the Greco-Persian Wars that the Greek artists began to see the merit of observing life and living forms and of adding a whole new dimension of expression to their storytelling.

Commerce. In Homeric Greece commerce was not an honorable profession, but, by the end of the Archaic period, the survival and wealth of most Greek city-states depended on trade, and commercial advantage could be a cause of war. The demand for raw materials, manufactured goods, and food in the expanding Greek economy was generally in this period met by private enterprise. A merchant embarked with his goods, paid his passage, and sold where he would; or, like Colaeus the Samian, it might be the captain of a ship who took what opportunities and profits were presented. Although it was to the state's advantage that there should be enough food or bronze for a hoplite army, it was some time before the state, as such, took a leading part in fostering or directing trade, although some tyrants did, it seems, as at Corinth. And, in early 6th-century Athens, Solon's reforms were in part directed at the Attic economy by encouraging an export market in oil. There was also an indirect interest in trade wherever taxes were levied on goods entering or leaving ports or passing through controlled territory. Slowly an ethic of commerce was evolved as trade became of more pressing importance to states. Piracy probably became less menacing, which is remarkable since there was no central government, let alone any policing force; it speaks well for the common sense of the Greeks and their readiness to submit dispute to the arbitration of fellow Greeks. It seems to have been remarkably easy for goods to pass between Greece and non-Greek enemy territories. There was a naïve simplicity about Archaic trade, but the days of tariff barriers and economic warfare were not far off.

Local or retail trade was, for most of the period, conducted by barter, the countryman exchanging his foodstuffs for manufactured goods or recognized precious objects, such as iron spits, which served as a form of primitive currency. But there was no question of the weighing of precious metals for retail payment, and for a long time Greeks were incapable of precision weighing. When dumps of precious metal were first stamped with a state or personal blazon (by about 600 BC, in Lydia, it seems) this guarantee was of the issuing authority rather than of weight. The use of coinage spread quickly through Greece. It was for local exchange, however, perhaps for payment of mercenaries by the state, and certainly for payment of fines or taxes. But there was seldom any provision of small change to serve a retail market and no passage of money from state to state to indicate its use for general trade. Only toward the end of the Archaic period and in places in which the precious metal was in good supply, as in Athens, was the production of coinage by the state apparently intended to serve "internationally" as recognized bullion. This is the beginning of a money economy, but coinage was not invented to serve this end, nor did it serve it for a century or more. (Jo.Bo.)

Introduction of coinage

THE GREEK CITY-STATE

The growth of the Polis. The period of nearly three centuries in Greece after the great migrations of *c.* 1100–*c.* 1000 BC is often called the Greek "dark ages" because little is known about it. Habitation was in villages and towns, which have left few traces except for pottery of Proto-Geometric and Geometric types. Rather before 700 BC a revival began, largely because of Oriental influence, and the term Archaic is given to the age *c.* 750–*c.* 500 BC to distinguish it from the Classical period that followed. Among the outstanding developments of the Archaic age was the growth of urban life and political institutions.

Character and early condition. The type of settlement emerging in many parts of Archaic Greece was the *polis* (plural *poleis*), or city-state. The *polis* was a sovereign political unit. Usually it had a fortified centre on a hill, where the inhabitants took refuge from attack; a residential and commercial town arose at the foot of the hill, and the town was surrounded by a dependent territory, which served for agriculture and pasture and, in the case of large *poleis,* included villages subordinate to the main city. Thus at Corinth the fortification was on the steep hill called Acrocorinth; the town of Corinth arose at the foot of the hill and commanded an extensive territory, which included subsidiary settlements such as Lechaeum, Sidus, and Crommyon. (It should be noted that this was not the only characteristic type of settlement; in much of European Greece, especially in the northwest, the unit was not the *polis* but the *ethnos* [plural *ethnē*], or nation. An *ethnos,* such as Phocis or Aetolia, held extensive land, which was divided among unfortified villages; the villages were linked in a loose federation, and no single centre predominated. But until the 3rd century BC, the political development of the *ethnē* was slow.)

In the early Archaic *polis,* public authority was weak. Society was agrarian, and each community had people of

quite varied status—wealthy landowners, small farmers, landless labourers, artisans, and in some places serfs. The strong unit was the family; a large and powerful family attracted numerous dependents, giving them protection and economic help in return for their labour.

The divisions in society, consequently, were not horizontal but vertical: conflicts were not between classes but between groups. Each group was led by one or a few strong households, which had many dependents. (These groups have been aptly compared to pyramids because each had a broad base of humble men.) Eventually these relationships became formalized and hereditary; the consequent terms are known in the case of Athens. There *genos* (plural *genē*), or clan, was the word for a group of powerful families that led a body of dependents; each *genos* claimed descent from a common ancestor, who was usually mythical. The *genos,* together with the body of dependents attached to it, constituted a phratry, or *phratria.*

The citizens of each *polis* were also divided into hereditary units called *phylai* (singular *phylē*), a word often translated by the Roman term tribes. In Dorian *poleis* there were usually three tribes—the Hylleis, the Pamphyli, and the Dymanatae; among the Ionians the number was less uniform, but some of the same names occur in several Ionian *poleis.* Evidently the tribes date back to the age of migration, before they had settled in cities.

Political institutions

Three political institutions were commonly found in each early *polis*—a king, a council, and an assembly of adult male citizens. These institutions arose understandably from the migrations. Each migrating host required a war leader, whose office developed into hereditary kingship when the host found land and settled down. The war leader consulted the leaders of distinct contingents (his council) among his followers before taking action; and he announced his decisions to a gathering of all his able-bodied men, assembled to receive marching orders. (Such practices are reflected in the *Iliad* and the *Odyssey.*) But in sedentary conditions changes came about. The kingship was more vulnerable than the council: the king might be a minor, or there might be a disputed succession. In many cities additional offices, often elective and annual, were created alongside the kingship; even the kingship became annual and elective in some places, including Athens.

The council, though at first merely advisory, gained in power from the decline of the kingship and became the primary scene of political rivalries. The assembly changed from a gathering that heard decisions to a gathering that reached decisions. This does not necessarily mean, as has often been supposed, that the common people, working through an assembly, asserted themselves against an aristocracy, working through a council. More likely, it means that when leaders of prominent families could not get their way in council, they took the issue to the assembly, calling out their followers to fight for and later to vote for them; and as this practice became habitual, the assembly became a second place for settling political rivalries.

Oriental influence. During the 8th century BC, Oriental influences began to reach the previously static societies of the Greek communities. A Semitic and probably Phoenician alphabet was borrowed and adapted for writing Greek. Perhaps in the same period Greeks learned the Oriental stories that appear in the *Theogony,* an account of the beginning of creation and the birth of the gods by the poet Hesiod (fl. *c.* 800 BC). Changes in pottery provide the clearest indication of foreign influences in Greek societies. Geometric pottery was superseded by new naturalistic styles in different cities at different times during the 8th and 7th centuries. Collectively the new styles are called "Orientalizing" styles, but they varied according to place of manufacture. The pottery made at Corinth was more widely exported than any other during the 7th century; the subdivisions of the proto-Corinthian and Corinthian styles span the period *c.* 725–*c.* 550. The adoption of the Orientalizing styles and the fact that these new styles travelled farther than had Geometric pottery indicate a growth of overseas trade, and this was likely to dislocate the traditional societies of the rudimentary Greek *poleis.* The colonization movement (*c.* 750–*c.* 500 BC) is evidence of profound social changes.

Hoplites. A change in military practice—the introduction of hoplite equipment and hoplite tactics—had effects that are difficult to estimate. In the "dark ages," fighting was mainly conducted by comparatively few warriors: each had a shield, a sword, and spears, usually for throwing, but lacked armour for the head and body; warriors fought as individual champions, not as an organized group. In the 6th and 5th centuries the characteristic force of the developed *poleis* was the phalanx of hoplites, or heavily armed infantrymen. The hoplite had a sword and a thrusting spear; and above all, he had plentiful protective armour, including a closed helmet, a bronze plate corslet, greaves, and a large round shield with armband and handgrip. In traditional practice the front row of hoplites tried to push the enemy off the field, and the succeeding rows of the phalanx added weight to the impact.

Until recently it was believed that many Greek states had adopted the hoplite phalanx within a few years of 700. But fresh study of vases that depict armed men has revealed a more complex development of the hoplite phalanx. The various items of hoplite equipment were adopted separately in the period 750–700, but at first the old manner of fighting by uncoordinated champions was not changed. The complete hoplite panoply is first attested on a vase *c.* 675, and the massed hoplite phalanx first appears *c.* 650. Moreover, the social effect of the change in tactics is disputable. A widely accepted theory says that in consequence of the change, the burden of fighting passed from relatively few aristocrats to a wider class of substantial peasant farmers; that the latter accordingly demanded a share in political power; and that revolutionary rulers, the tyrants of the 7th century, arose as their leaders. Against this view it has been argued that peasant farmers were more likely to be conservative than revolutionary, and they would dislike war and disturbance as threatening their land and taking them away from tillage; in Etruria and Rome hoplite tactics were adopted during a period of aristocratic ascendancy. Perhaps the social significance of the hoplite phalanx was something different: the new type of fighting was much more highly organized than was the old, and so hoplite discipline contributed to the growth in the authority of state.

Tyrannies in Archaic Greece. Some increase in public authority was brought about by the tyrants, who arose in many cities—especially of the Peloponnese—during the 7th and 6th centuries BC. The words tyrant and tyranny do not necessarily imply oppressive rule. Essentially a tyranny began when a man without any hereditary or official claim to rule seized control of his city. Usually a tyrant held no office and did not alter the city's laws or institutions. He exercised an informal and extralegal preeminence; the nearest modern equivalent is the "boss" of a small town.

The rise of tyrannies

Because several tyrannies began in the Peloponnese within a limited period, scholars have often sought a common cause for them all. Some scholars have suggested commercial factors; others have argued that tyrants arose as champions of the pre-Dorian element in the population against Dorian aristocracies; others have attributed their rise to class conflicts consequent upon the introduction of hoplite tactics. But no one factor can be adequately attested for all the tyrannies of the Archaic Peloponnese. Local conditions varied, and the only note common to all tyrants was one of magnificence; they made a display of their power and wealth.

The word tyrant was not Greek in origin; it may have been a Lydian word for "king," and the Lydian kings of the Mermnad dynasty (*c.* 680–*c.* 544 BC) impressed the Greek imagination with their power and display. In Greek the word tyranny first occurs (7th century) in a quatrain of the poet Archilochus, who there associates it with his contemporary, Gyges, the founder of the Mermnad dynasty, and with ostentatious wealth.

Pheidon, a hereditary king of Argos, was a forerunner of the tyrants. Aristotle says, enigmatically, that he began as king but became a tyrant. He brought the whole of the Argolid (the area of Argolis) under the control of Argos; this was the basis for further expansion. Pheidon proceeded southward into the plain of Thyrea, which the



Greek expansion *c.* 550 BC.

Adapted from W. Shepherd, *Historical Atlas;* Barnes and Noble, Inc.

Spartans too wished to seize; the Argives defeated them at the Battle of Hysiae in 669, and this ensured control of Thyrea to Argos for more than a century.

A tradition known to Herodotus said that Argos once ruled the whole southeastern coast of the Peloponnese as far as Cape Malea and even the island of Cythera. This tradition, if sound, must refer to the time of Pheidon, but it may be exaggerated. Probably in 668, although the date is disputed, Pheidon intervened at the Olympic festival; he excluded the regular Elean presidents of the competition and took the presidency himself. In this action he took advantage of tension within the territory of Elis. There the Eleans, who had arrived during the migrations, had ascendancy; but there was a rural and underprivileged class, the Pisatans, who were descended from those who had inhabited the land before the migrations. Pheidon promoted the aspirations of the Pisatans, and from his time intermittently until 580 they succeeded in seizing the presidency at the Olympic competition. Pheidon evidently achieved military ascendancy in much of the Peloponnese. He struck standard weights and measures, dedicating the models at the Argive Temple of Hera, and these standards were accepted permanently by many Peloponnesian states.

The fullest information on early tyranny concerns Corinth and its western neighbour Sicyon. Under Cleisthenes (reigned *c.* 600–*c.* 570 BC), Sicyon achieved greater power than at any other time. Herodotus, the best source, tells two stories about Cleisthenes of Sicyon. In one he relates how Cleisthenes, being at war with Argos, took several anti-Argive steps in domestic affairs; for example, he tried to suppress the cult of the Argive hero Adrastus. Further, according to the story, Cleisthenes altered the names of the tribes at Sicyon. Previously there were the three Dorian tribes of Hylleis, Pamphyli, and Dymanatae, and a fourth tribe called Aegialeis; Cleisthenes gave his own tribe the name Archelai (Rulers of the People), but he renamed the Dorian tribes, calling them Hyatae (Pigites), Oneatae (Assites), and Choereatae (Swine-ites). Thus Herodotus represents the tribal change as merely a change of names carried out because of the war against Argos; Cleisthenes did not want the Sicyonians to have the Dorian tribes in common with the Argives. But something more serious must have been intended. By Herodotus' time the tradition was already becoming anecdotal; the story of the change in names conceals a real reform. Previously, the non-Dorian element in the population—that is, the descendants of those inhabitants who held Sicyon before the Dorian settlers arrived—were underprivileged, at least to the extent of constituting only one tribe in contrast to the Dorian three. Cleisthenes altered this condition drastically; the change was in the nature of a social revolution. Doubtless it was carried out not because of the war with Argos but for internal reasons. Moreover, the conjecture can hardly be avoided that this social reform was the source of Cleisthenes' power: he ruled as the champion of the pre-Dorian element in the population.

Reforms of Cleisthenes

The other Herodotean story tells how Cleisthenes sought a bridegroom for his daughter, Agariste. On winning a chariot victory at Olympia, he invited prospective suitors to Sicyon. Men came not only from the Peloponnese but from places as far away as Crannon in Thessaly and Sybaris and Siris in southern Italy. Cleisthenes inquired into their descent and kept them at his court for a year, testing them in all manner of gentlemanly exercises. He preferred the two Athenian suitors, Hippocleides and Megacles, especially the former. But on the last evening, after dinner, Hippocleides climbed on a table, stood on his head, and danced with his legs in the air. When Cleisthenes told him that he had danced away his marriage, Hippocleides retorted that he did not care. So Cleisthenes chose Megacles; he said that he honoured them all and would gladly gratify them all, but with only one daughter to give away he could not please everyone.

This story may have gained in the telling, but it reveals much. Hippocleides of the Philaid clan traced his descent to the legendary hero Ajax. Later writers said that Cleisthenes was the great-grandson of a butcher or cook; Herodotus made the point more effectively by putting a banal speech into his mouth. By inviting suitors to compete for his daughter, Cleisthenes made a display of magnificence. He had to; he was a self-made man.

Writers later than Herodotus made Cleisthenes a member of a dynasty, which took its name from its founder, Orthagoras, and which allegedly ruled Sicyon for 100 years;

the dynasty may have lasted from *c.* 656 to *c.* 556, but modern research has not fully determined the chronology. Even if Cleisthenes had predecessors in the tyranny, he transformed its basis by the tribal reform. Perhaps the only fact of value provided by the later sources is that Cleisthenes joined in the First Sacred War against Crisa (see below *Relations among the Greek poleis to the end of the 6th century BC*), which stood on the north shore of the Corinthian Gulf. Sicyon on the south shore could regard Crisa as a commercial rival, and trade was the ultimate source of Sicyonian prosperity under the tyranny.

The prosperity of Corinth, too, depended on overseas trade. In the 8th and early 7th centuries the city was ruled by a group called the Bacchiads, whose members traced descent from a legendary common ancestor and were so exclusive as to practice intermarriage within the group. But one of their daughters, Labda, was lame and thus was given in marriage to Aetion, who was not one of the Bacchiads; he claimed descent from the pre-Hellenic inhabitants of Greece and came from Petra, a village in Corinthian territory at some distance from the city. Their son was Cypselus; and when he came of age, he overthrew the Bacchiads and started the Cypselid tyranny of Corinth. He killed some of his opponents and exiled others. After ruling for 30 years he was succeeded by his son Periander, who continued and intensified his father's severities; Periander kept a bodyguard of spearmen, whereas none is attested for Cypselus. Periander was succeeded by his nephew Psammetichus, who was overthrown after a short reign. The Cypselid tyranny is commonly dated *c.* 657–*c.* 584 BC, although somewhat later dates (*c.* 610–*c.* 540 BC) can be defended.

The tradition recorded by Herodotus and Aristotle has not been kind to Cypselus or his dynasty; it stresses their overthrow of the Bacchiads and their alleged cruelties. Yet a comparison of Cypselid policy with Bacchiad policy reveals not contrast but continuity. The export of Corinthian pottery flourished increasingly until *c.* 575; then the ware began to decline in quality, and after *c.* 550 it yielded to Attic competition, which may have had to do with the fall of the tyranny.

Corinthian colonies

Corinthian trade was helped by colonization. Under the Bacchiads, Corinthian colonies were established at Syracuse and Corcyra *c.* 733. Cypselus founded colonies at Leukas, Anactorium, and Ambracia; these commanded the Gulf of Ambracia and the approach to southern Epirus. Early in the time of Periander, the Corcyreans sent a colony northward to settle at Epidamnus on the coast of the Adriatic; they invited Corinth to send a leader of the colony, and the man sent was a Bacchiad. (This apparently indicates a reconciliation between the Cypselids and at least some of the Bacchiads.) Indeed, Periander controlled Corcyra and eventually sent one of his sons to rule it. A Corinthian colony was founded at Apollonia, somewhat south of Epidamnus, probably in the time of Periander; Epidamnus and Apollonia were situated at the two western termini of one of the few natural routes leading into the Balkan highlands and even into Macedonia. A son of Periander led a colony to Potidaea on the Isthmus of Pallene in the northeast; this, too, was well placed for opening up Macedonia.

Besides colonization, other items of Cypselid policy assisted trade. The tyrants, for example, built a stone causeway, or *diolkos,* across the Isthmus of Corinth. Commerce may have been among the considerations inducing Periander to maintain good relations with Lydia and Egypt.

Little is recorded of other contemporary tyrants in or near the Peloponnese. Theagenes ruled Megara *c.* 632. According to Aristotle, he caught the flocks of the wealthy at pasture and slaughtered them; this may mark the overthrow of a wool-growing aristocracy. Procles, the tyrant of Epidaurus, is known only from his relations with Periander. The latter married Procles' daughter, but after her death he quarrelled with his father-in-law and overthrew him.

Conditions producing the tyrannies

The conditions producing the several tyrannies varied; no one common cause can be specified. Tension between the Dorian and pre-Dorian elements in the population, for example, was important in Sicyon and in Elis, but it has not been proved for Corinth. Admittedly, in virtue of his paternal descent Cypselus appears to have begun as a man on the fringe of the ruling group, and he may consequently have been motivated by jealous ambition; but his policy, so far as known, did not attempt any drastic change in the social structure of his city. Yet the note of magnificence, common to the different tyrants, may have had political significance. They set a new standard of wealth and power; the first stone buildings in Greece since the Bronze Age were the work of the tyrants. (For example, the Corinthian treasury at Delphi was built by Cypselus.) Such achievements marked a greater concentration of resources and authority than had been known before. It is often said that the tyrants overthrew traditional aristocracies; and this is not incorrect, but it can be misleading since "aristocracy" may suggest a regime within a fully organized city. In fact, under the traditional aristocracies the organization of the *polis* was still rudimentary; the Archaic tyrannies marked a stage in the growth of public power.

Sparta. The territory south of the Arcadian highlands, which form the centre of the Peloponnese, consists of two plains bounded by mountain ranges. In ancient times, the eastern plain was Laconia and the western plain was Messenia. The plain of Laconia was bounded by Mt. Taygetus on the west and Mt. Parnon on the east, and it was drained by the Eurotas River. Dorian invaders reached Laconia in the age of migrations; they came as conquerors but mingled to some extent with the previous inhabitants. The earliest Laconian pottery of Proto-Geometric type belongs to the period *c.* 1000–*c.* 950 BC; it may mark the beginning of settled conditions, as a number of townships arose in the plain. One of these, Sparta, situated well up the valley of the Eurotas, consisted of four villages—Pitane, Mesoa, Limnae, and Conoura. The account given by Pausanias of the early expansion of Sparta is credible, although one cannot be sure whether it rests on genuine tradition or on plausible reconstruction. According to this account, the Spartans first moved northward, conquering Aegys at the approach to Arcadia. In the next generation they advanced southward against the strong town of Amyclae—a long war that led to the absorption of that town into the city of Sparta, which thenceforth consisted of five units instead of the original four—and against the towns of Pharis and Geronthrae, whose inhabitants were expelled. Pausanias' account continues by saying that, in the generation following these conquests, the Spartans overcame Helus on the coast; other coastal towns, such as Las and Gythium, probably fell about the same time. These gains should belong to the 9th and 8th centuries; they built up a large state, embracing the plain of Laconia.

Early expansion of Sparta

The Messenian Wars. The population of Greece was growing, and many states sought relief in colonization. Sparta founded only one colony, namely Tarentum in southern Italy; its earliest pottery belongs to the period *c.* 700–*c.* 650 BC. For the most part, Sparta sought land for its surplus population by expanding against its neighbours. Expansion was attempted in two directions. Northeast of Mt. Parnon lay the coastal plain of Thyrea; there Spartan ambitions came into conflict with those of Argos. In 669 the Argives defeated the Spartans at the Battle of Hysiae; this had the effect of excluding Spartan interests from the plain of Thyrea for more than a century.

The other direction of expansion led westward into Messenia. This plain had been settled by Dorians, and numerous townships had arisen there, probably in much the same way as in Laconia. A Spartan attack on Messenia led to protracted warfare. The poet Tyrtaeus, exhorting the Spartans at a late stage in the struggle, said that under King Theopompus "the fathers of our fathers" conquered Messenia; he added that they fought there for 19 years but in the 20th year the defenders fled from the mountains.

Conquest of Messenia

The chronology of the fighting is obscure. After 370–369 BC, when the Theban commander Epaminondas led a successful campaign against the Spartans and restored Messenian independence, a previously sparse tradition on early Messenian history became contaminated with romantic inventions. Many modern historians believe that there were two early Messenian Wars: the first (*c.* 735–*c.* 715), was the first Spartan conquest of Messenia; but

a Messenian revolt (*c.* 660) precipitated a second war, in which the Spartans were ultimately successful. The date suggested for the first war rests largely on an argument from the list of Olympic victors: until 736 the victors are from the western Peloponnese and chiefly from Messenia; but after 736 there is no Messenian, and the first Spartan victor is recorded in 716. But the reliability of the early entries in the Olympic list is an unsolved problem. Moreover, it must be admitted that although writers after 369 distinguish between two early Messenian Wars, earlier writers such as Thucydides (writing in the late 5th century) merely mention fighting without specifying any number of wars. Perhaps the most valuable indication of date is a remark of the Theban general Epaminondas; in 369 he said that he had restored Messenia after 230 years. Evidently something concerning Messenia came to an end about 600.

The conquest of Messenia, completed by the end of the 7th century BC, gave the Spartans extensive land. Much of it was divided into allotments for Spartans. The whole state, consisting territorially of Laconia and Messenia, was properly called Lacedaemon; it was a federal state of somewhat authoritarian type. The institutions of the city of Sparta had sole direction of foreign policy and exercised supervision of all Lacedaemon.

Among the inhabitants of Lacedaemon, three main statuses should be distinguished. First there were Spartan citizens—the most privileged group. Each family of Spartan citizens had an allotment (*klēros*) of land in Laconia or Messenia; the allotments were deemed to be equal. Secondly there were the *peroikoi,* or free inhabitants of the other towns of Laconia and of those Messenian towns that were allowed to survive after the conquest; each perioecid town had some local autonomy. Finally there were the helots, whose name may be derived from the town of Helus. They were serfs bound permanently to the soil; a group of them was assigned to each Spartan allotment of land, and they tilled it for the Spartan masters as well as for their own subsistence.

The new political system. Aristotle, referring to Tyrtaeus as his source, says that during the Messenian War some people in Lacedaemon demanded a redistribution of land. The protracted warfare against the Messenians and the eventual conquest of their land probably caused extensive political changes within Sparta, but the nature of the change is obscure. Here it will be convenient to consider first the institutions of the city of Sparta as they were in the 6th century BC, and then to try to discern the changes that produced them.

The Spartans attributed their institutions to an early lawgiver called Lycurgus; scholars, however, have not determined whether he was real or, if real, what institutions should be attached to his name.

The dual kingship

Sparta had two hereditary kings, drawn from two royal families. It is often conjectured that the dual kingship arose from the early amalgamation of different communities, perhaps from some combination of the original four villages. The powers of each king included command in warfare, and during the Archaic period this carried with it a large share in determining foreign policy. But by the later decades of the 5th century, decisions on foreign policy were taken in the public assembly; and in their military capacities the kings became merely the executors of the assembly's will, although they retained discretionary control of strategy in the field. In home affairs the kings enjoyed some priesthoods and perquisites, they were members of the council, and they exercised judicial powers of three kinds: they judged cases concerning public roads, they arranged the marriages of unbetrothed heiresses, and they presided over all adoptions of sons within the community. These functions may at first seem slight, but in Archaic societies, where public authority was weak and the family was the strong unit of society, matters concerning heiresses and adoptions were important. Later the domestic powers of the kings did not decline, but the state undertook additional functions through other organs.

The council, or *gerousia* (*gerōn,* "elder"), consisted of 30 members including the kings. The other 28 councillors had to be at least 60 years old; they were chosen by acclamation in the public assembly and remained in office for the rest of their lives. The council could advise the kings, served as a court to try capital cases, and probably had the task of preparing business for the public assembly. The extent of its real influence in politics is a delicate question and requires careful attention.

Any adult male citizen of Sparta could attend the public assembly. In addition to electing councillors, this body elected officers called ephors and had the ultimate decision on questions of legislation and policy. Voting was usually by acclamation, a procedure that allowed some discretion to the presiding officer, but occasionally a division was taken.

A king, a council, and an assembly were the usual pattern in any Greek state; but Sparta developed a further major organ—the ephorate. The ephors were five in number and were chosen by acclamation in the public assembly; each group of five held office for only one year. In the 5th and 4th centuries the ephors transacted a good deal of day-to-day business. They presided at meetings of the assembly, and, because they formulated an issue for voting, they could sometimes influence the outcome. Probably they could preside at meetings of the council. They carried out the instructions issued by the assembly for beginning military campaigns, and they transmitted orders to commanders in the field. They received foreign envoys and provided for bringing their business before the assembly. They exercised supervision over the helots.

The ephorate

The ephorate may have existed in some form from very early times. Each month the ephors exchanged oaths with the kings; the kings swore to rule according to the laws; and the ephors, on behalf of the city, swore to preserve the kingship provided that the kings kept their oaths. Since the oath was monthly, not annual, it may have come from a time when the moon, but not the sun, was observed for calendric purposes. But the functions of the ephors were probably small at first and grew later. Spartan tradition held that the powers of the ephorate were much increased by Chilon, who was ephor for a year within the period 556–553, but nothing specific can be said about his work in internal affairs. Scholars once supposed that for periods of many years the ephors pursued a single corporate policy in opposition to the kings, but this is not plausible; because there were five ephors in office together and because they changed each year, it is more likely that rival political groups used every opportunity to get their own sympathizers elected to the ephorate. Indeed, instances are known where one or two ephors pursued a policy distinct from that of their colleagues and their immediate predecessors. To explain the growth in the significance of the ephorate, perhaps the most likely hypothesis is that public authority, as it grew stronger, increasingly assigned new functions to the ephors.

A special question concerns the relationship between the ephors and the kings. By the time of Thucydides, the ephors could imprison the kings; but this was merely a power of arrest with a view to trial, not a power to judge and condemn. An incident occurring about the mid-6th century is suggestive. King Anaxandrides had a wife but no children, so the ephors urged him to divorce her and take another wife in her place. He refused. As a next step the ephors and the councillors together urged Anaxandrides to take a second wife in addition to the first, and they suggested that the Spartan citizens might take action against him if he did not comply. Although he did comply, it is not clear whether he yielded to the authority of the councillors or to the suggested threat from Spartan citizens. But it is evident from the incident that the ephors alone could not coerce the king.

Such were the main institutions of Sparta in and after the 6th century. Discussion of the way they developed in the previous century has concerned itself mainly with the relationship between the council and the assembly. Plutarch in his *Life of Lycurgus* preserves a document that may have a bearing on this. Although its text is uncertain at some points, probable readings give the following sense:

> After the people has set up a sanctuary of Zeus Syllanios and Athena Syllania, after the people has arranged itself by tribes and obes, after the people has set up a council of thirty together with the kings, let them gather from season to season

for the festival of Apellae between Babyka and Knakion; let the elders introduce proposals and decline to introduce proposals; but let the people have the final decision.

Commenting on the text, Plutarch says that two later kings, Polydorus and Theopompus, added a further clause; this addition, now often called "the rider," says: "But if the people make a crooked utterance, let the elders and the kings decline it."

The Spartan *rhētra*

Plutarch says that the document was an oracle, which Lycurgus brought from Delphi, and that the Spartans called it a *rhētra.* The word *rhētra* means here "an enactment"; and the name suggests that the document was an enactment of the Spartan assembly, even if it was based on an oracular response. Its authenticity can be disputed on linguistic grounds. Even if it is authentic, it mentions a good many institutions that required further definition; so it can only have been part of a body of legislation or perhaps a summary of measures, which required further and more specific enactments.

The main clauses of the *rhētra* deal with the decision-making functions of the council and the assembly, the festival of Apellai being the occasion for meetings of the assembly. The *rhētra* provides that measures are to be introduced by the council but that the final decision shall be taken by the assembly, which thus has at least a yea or nay competence. A two-stage procedure of this kind is attested later in Athens and in other Greek states.

According to a theory accepted by many present-day historians, the two-stage procedure was introduced by the *rhētra* and was a Spartan invention of the 7th century; it appeased strife between the aristocracy, whose organ was the council, and the wider class attending the assembly; and it achieved this by defining their respective shares in determining policy. The theory has also a social aspect. The hoplite phalanx is first attested *c.* 650, and adoption of hoplite tactics may have enabled the Spartans finally to overcome the Messenians; if their adoption spread the burden of fighting to a wider class than before, this may explain why that class demanded a share in making policy.

This theory is impressive. It is less than explicit, however, about the change made by the *rhētra* and about what the previous condition was. A public assembly of some kind must have existed since the age of the migrations, when it began as a parade of the army to receive marching orders; analogies with other Greek states suggest that in settled conditions such an assembly is never likely to have had less than a yea or nay competence. Moreover, a two-stage procedure for determining policy was not necessarily a deliberate Spartan invention adopted subsequently by other Greek states. Comparable procedures are attested in the Roman Republic—where it was customary to bring measures to the Senate before presenting them to the assembly—and among the ancient Germans. Indeed, in the *Iliad,* Agamemnon, having conceived a plan for withdrawal, first discusses it with a "council of elders" and then summons the army to hear it. From these parallels it appears that in Sparta and in other Greek states the two-stage procedure may have arisen from custom and convenience, without any specific enactment designed to appease a conflict.

The problem is complicated by a fragment of a poem of Tyrtaeus. It reads thus:

> After listening to Phoebus they brought home from Pytho the oracles of the god and perfect words: "Let the lead in deliberation be taken by the god-honored kings, who have care of the lovely city of Sparta, and by the elderly councillors, and then let the men of the people, answering with straight enactments. . . .

The fragment breaks off. It can be read as a paraphrase of Plutarch's *rhētra,* but even so it is much more monarchical in spirit. Alternatively, Tyrtaeus' remarks can be adequately understood even if the *rhētra* is inauthentic; in that case he would seek to commend the two-stage procedure in a monarchical spirit by alleging that it was sanctioned by an oracle. Indeed the fragment is good evidence for only one assertion, namely that the Spartans followed the two-stage procedure.

Plutarch introduces the *rhētra* to indicate the importance that Lycurgus attached to the council. If the rider is combined with the *rhētra,* the result gives the council unusual authority, for the rider grants the council some power of veto over the assembly. Obscure passages of Aristotle and Diodorus have been used to explain and illustrate the provision embodied in the rider, but their proper interpretation is not wholly clear. Until recently it was commonly believed that in Sparta the real policy decisions were usually taken in the council; the assembly, it was supposed, failed to assert itself and merely ratified decisions of the council. New research has led to a different conclusion. The best evidence on Spartan procedure concerns the 5th and 4th centuries and is provided by Thucydides and Xenophon; these two authors hardly ever mention the council, but they indicate regularly that real decisions on policy were taken in the assembly.

Discussion of 7th-century changes in Sparta must be largely inconclusive, but some probabilities can be indicated. It is likely that the struggle with Messenia caused internal tension. It is also likely that hoplite tactics were adopted by the Spartans during the 7th century and had political effects. Those effects need not have been an exacerbation of tension; on the contrary, hoplite discipline may have accustomed the citizens to accepting a higher degree of organization. By the early 6th century, Spartan institutions were relatively articulate and defined; a hundred years earlier they had probably been much more amorphous and fluid.

The *agōgē*

Social organization. Sparta developed for its citizens an elaborate system of public education and discipline, often called the *agōgē.* At the age of seven, children were taken from their mothers and began to pass through a series of annual stages. The content of education included music and poetry as well as physical exercises and games. Girls were freed from public control when they reached maturity, but for boys, education culminated in two years of intensive military training. Then at the age of 20 each man sought admission to one of the living and eating groups, called *phiditia.* Each *phidition* had 15 members and served as a unit on military campaigns; each member contributed to the common stock a fixed monthly amount of food drawn from his *klēros,* or land allotment. The members of each *phidition* lived together until they reached the age of 30, when they could live at home with their wives. For military purposes men were summoned according to year classes, and the first incidence of fighting fell on those between the ages of 20 and 30; but they could be required for service until they were 60 years old.

The *agōgē* probably takes its origin from changes carried out in the 7th or early 6th century BC. It included initiation ceremonies that may go back to an earlier date; what calls for explanation, however, is not the details of the *agōgē* but its incorporation into a well-designed system. As that system trained men for fighting as hoplites, it is not likely to have originated before the adoption of hoplite tactics. The success of the system is reflected in Sparta's military victories from the mid-6th century onward.

The adoption of the *agōgē* may be connected with the introduction of a new way of organizing the citizen body. Originally the Spartans were divided into the three Dorian tribes, membership being hereditary. Tyrtaeus mentions the three tribes in a passage that seems to be an exhortation to Spartans of his own 7th century. But eventually a different method of division, on a territorial basis, was adopted. Spartan inscriptions of Roman date attest citizen units called obes; and the names of these coincide for the most part with those of the five divisions of classical Sparta—with Amyclae and the four original villages. Further, Aristotle says that Sparta had five ancestral regiments; he names them, and one of the names—Mesoa—is that of one of the villages.

Territorial division of citizens

These facts led to the hypothesis that the division into three tribes was replaced by a division into five territorial units called obes. Because membership of the tribes was hereditary, they would have become unequal in size; parallels in other states suggest that the aim in adopting the new division was to distribute evenly the burdens of citizenship, including military service. The problem became complicated by recognition of an Archaic Spartan inscription that names a sixth obe, called Arkaloi. Theo-

ries have been propounded that combine the division into three hereditary tribes with a division into numerous territorial obes, but such schemes would leave the ultimate units unequal and thus frustrate the purpose of introducing the territorial division of citizens. The figure five was significant in Spartan institutions; both the ephorate and a board of public messengers called *agathoergoi* consisted of five men each year. This fact and Aristotle's reference to five ancestral regiments suggest that the territorial division of the citizens was fivefold; whether "obe" was the word for the five chief divisions or for subdivisions of them is another question.

The social structure of Sparta produced a strong military force that served both to maintain the city's ascendancy in Lacedaemon and to expand its power in the Peloponnese. It does not follow, as has sometimes been asserted, that life in Lacedaemon became narrow, militaristic, and inartistic as early as the 6th century. Pindar praised Sparta as the place "Where the counsels of the old men and the spears of the young excel, as do dances and the Muse and splendor." The crafts and trade were conducted mainly by the *perioikoi* because Spartans were excluded from trade. Although the use of coinage was not adopted in Lacedaemon, archaeological finds show that the export of Laconian vases and bronzes continued through the 6th century. Laconian vase painting, though not among the best in Greece, was respectable and reached its highest level *c.* 560–550; its subsequent decline was not as acute as in the case of Corinthian pottery and can best be explained by the competition of Attic ware. Laconian craftsmen continued producing bronze figurines of good quality into the first half of the 5th century.

Life in Lacedaemon

The condition of the helots was severe, but their misfortunes and discontent in the early period should not be exaggerated. Tyrtaeus indeed compares them to asses worn down by heavy burdens and says that they surrendered half of the produce of the soil. But he wrote during warfare, and "half" may be approximate; Plutarch says that each *klēros* had to supply its master with a fixed and uniform amount of produce, and this suggests that the helots could keep the surplus. For the Battle of Plataea in 479, Sparta supplied 5,000 citizen hoplites and 35,000 helots; the helots were servicing troops, not fighting troops, but it would have been foolish to surround every armed man with seven thoroughly disaffected men, even if the latter were unarmed. Probably relations between Spartans and helots grew much more tense after the helot revolt of the 460s.

The Peloponnesian League. After the reduction of Messenia, and perhaps in consequence of it, Sparta came into conflict with some cities of Arcadia. The fighting culminated in a long war (*c.* 580–*c.* 550 BC) against Tegea. Although the Spartans may have gained some superiority, they were not able to overwhelm the Tegeans, and instead they concluded the war by making an alliance with their enemy. Shortly afterward (*c.* 544) the Spartans resumed their former struggle against Argos; the immediate issue was control of the plain of Thyrea, and this time the Spartans won a decisive victory.

The alliance with Tegea began a new development. In the 7th century the Spartan policy toward Messenia had been one of conquest and absorption; in the second half of the 6th century, Sparta tried to bring more and more cities of the Peloponnese into an alliance. Moreover, on allying with Tegea, Sparta brought from there the cult of the pre-Dorian hero Orestes; although Sparta remained a Dorian state, by this and subsequent actions it posed as the champion of the pre-Dorian element in the population, thus exploiting an issue that some of the tyrants had used. After defeating Argos, Sparta was evidently the strongest state in the Peloponnese, and during the next decades it succeeded in winning most of the Peloponnesian cities as allies. The resulting organization is called the Peloponnesian League. The stages of its growth are obscure; in 510 the Spartan king Cleomenes I led a force by land against Athens, and this shows that he had the goodwill of Corinth and Megara.

Spartan alliances in the Peloponnese

The name Peloponnesian League is modern; the ancient name was "the Lacedaemonians and their allies." The league began from bilateral treaties between Sparta and the several allies. Eventually a congress of the allies met and contracted a multilateral agreement; this may have occurred in the closing years of the 6th century, when Spartan attempts to intervene in Athens precipitated a crisis among the allies. The main provision of the multilateral agreement was that all members would be bound by a majority vote of the allies. The league lasted till 366, although its extent varied from time to time. It was a new experiment in interstate organization. Regional leagues elsewhere (*e.g.,* in Aetolia and Boeotia) relied on a fiction of common descent; the Peloponnesian League was an artificial creation of political power.

Athens. *Early period.* The city of Athens, with its Acropolis, was one of the few Greek sites where habitation was not interrupted at the end of the Bronze Age. Indeed, Athens was comparatively prosperous in the Proto-Geometric and Geometric periods. But it is not proven that any political institutions survived from the Bronze Age. Moreover, it is most unlikely that Athens retained in the "dark ages" any ascendancy over other parts of Attica that it may have previously exercised. Attica, Athens' territory in the Classical period, was large, but its unification was only completed at a late date.

Geographically, Attica consisted of several plains divided by hills. The central plain held the city of Athens and a strip of coast, including the bay of Phaleron. Farther west was the plain of Thria, on which the largest settlement was Eleusis. Eastern Attica consisted partly of low-lying territory and partly of hills; and it held several settlements, the most important being at Brauron and Marathon.

A lengthy process of Attic unification was probably composed of two stages. In the first, the strongest settlement in each plain acquired control over the other towns of its plain; thus three or four powerful states emerged. In the second stage, Athens, which already controlled the central plain, absorbed the communities of the outlying districts. This second stage was completed not later than the first half of the 7th century BC, but perhaps not much before.

Unification of Attica

The result was a unitary state; the political organs of the city of Athens were the sole focus of Attic political activity. But in the 7th century, public authority was still weak; real power belonged to locally influential families, each commanding numerous dependents. Recollections survived of conditions that prevailed before the final steps of unification; late in the 5th century, people remembered that there had once been warfare between Eleusis and Athens. Accordingly, at least some of the political conflicts of the 7th and 6th centuries were struggles of "regionalist" type—that is, struggles of leading families in the outer districts against the ascendancy of those in the central plain.

The chief political institutions of the unitary state were an executive board called the "nine archons," a council, and a public assembly. Tradition held, credibly enough, that Athens had once been ruled by kings who retained office for life but whose powers had been diminished in two ways: other offices were created to take over some of their functions, and the term of all the officers was reduced to one year. Hence, the word king survived as the title of one of the nine archons; the others were the archon (or eponymous archon, since his name was used to designate the year), the *polemarchos* ("war commander") and the six *thesmothetai* ("setters of verdicts"). The evidence conflicts on the exact mode of selecting the nine archons, but it is safe to say that until 487 they were always elected by the assembly.

Athenian political institutions

The council came to be called the "council of the Areopagus" because it held some of its meetings on that hill west of the Acropolis. The council consisted of all who had held any of the nine archonships. Doubtless it had begun as a council advising the kings. (The public assembly, too, probably existed in some form from time immemorial, and questions were brought before it for ultimate decision; but probably its meetings were infrequent in the Archaic period.) Athenian writers of the 5th and 4th centuries BC knew that the council had once been powerful, but their attempts to define its former functions produced little more than vague generalities. Probably its ascendancy in the Archaic period rested not on specific powers but on a

personal factor. The minimum age for the nine archons was 30 years. As one of the archons, a man had only one year in office, but thereafter he could expect some 30 years of active life as a councillor. The archons were likely to respect the opinion of their seniors, to whose august company they would soon be admitted.

Cylon's uprising

Probably in 632 an uprising led by Cylon tested the stability of the unified state. Cylon had married the daughter of Theagenes, tyrant of Megara. Encouraged by an oracle, Cylon gathered his friends and a force from his father-in-law and seized the Acropolis during the Olympic festival. But the Athenians came en masse from the fields, beseiged the Acropolis, and entrusted operations to the nine archons. Cylon escaped but his followers ran short of food and made terms with the archons, who agreed to spare their lives; then, as the Cylonians came down from the Acropolis, the archons killed them. Later those responsible for the breach of faith were exiled as being accursed, but later still they were allowed to return. The scandal of the curse was revived more than once against their descendants.

The nature of the rising is obscure. Evidently Cylon hoped to make himself tyrant, but attempts to regard him as a champion of popular discontent have no evidential support. A clue may perhaps be found in the identity of his opponents. Whenever the scandal of the curse was revived, it was used against the powerful family of the Alcmaeonids and their associates; apparently the Alcmaeonids predominated among the archons of 632. As will be seen later (see below *Peisistratus and his sons*), the Alcmaeonids belonged to the central plain or possibly to southeastern Attica. Cylon could get forces without impediment from Megara; and this suggests, though by no means proves, that he belonged to western Attica. Possibly his rising reflects a conflict between families powerful in different regions.

Laws of Draco

Still greater mystery surrounds the activities attributed to Draco (Dracon) in 621. Later Athenian tradition said that he wrote down all the laws but that subsequently Solon repealed the laws of Draco except those on homicide. This story sounds artificial, and it is more likely that Draco concerned himself only with homicide laws. In 409 the homicide laws of Draco were reinscribed on stone, and the part of the inscribed text dealing with unintentional homicide has been recovered. The law provided that if a man killed another man unintentionally his case should be judged by a "Court of Fifty-One," called the *ephetai;* if they upheld his plea, the relatives of the victim could grant him pardon by a unanimous vote, but if the relatives declined to do so, the state provided the killer with a safe-conduct to the border.

Supplemented by evidence from the 4th-century orators, the inscribed text throws a good deal of light on early Attic procedure. The blood feud is implied, and the complementary institution of wergild (payment to relatives of a slain person, according to a determined value for the deceased) is attested. By recognizing the killer's intention as distinct from his act, the law moved toward greater sophistication; by setting up a court to judge intention and by providing a safe-conduct to the border for the unintentional but unpardoned killer, the state asserted its authority to a modest but real degree. Unfortunately the inscribed text of Draco's laws is incomplete.

By the 4th century, Athens had no less than five homicide courts, including the Areopagite council, which dealt with intentional homicide. This system of courts observed other distinctions besides those based on intention; for example, it recognized a category of lawful homicide, and it treated the killer of a slave or foreigner differently from the killer of a citizen. So complex a system must have taken a long time to develop, even though all the provisions may have had an early origin. It is not clear whether all parts of the system were attributed to Draco, and it is not clear which items were innovations of 621. But at least it can be said that by 621 the state had asserted itself in the originally private sphere of killing and the blood feud.

The reforms of Solon. Athens in the Geometric period was prosperous, but in the 7th century BC its economy was comparatively backward. Its failure to found any colonies in that century indicates political weakness at a time when population was doubtless growing in Attica as in the rest of Greece.

Early in the 6th century, Solon—in political poems of which parts are preserved—spoke of acute antagonism between rich and poor. This economic crisis has often been regarded as the result of impoverishment, but the archaeological record suggests a more complex explanation. The export of Attic pottery decorated with black figures on a red background began in the period 620–600, and the ware reached farther afield until *c.* 520; thereafter Attic ware retained the predominance it had achieved, but pottery decorated with red figures on a black background eventually became more popular than the black-figure ware. This export trade was not only in vases but also in their contents, perhaps wine and olive oil. Thus the Athenian economy was already expanding in the early 6th century. The expansion may have brought complex dislocation; in particular, amid growing prosperity, many people may have come to resent conditions they had previously tolerated.

Solon was eponymous archon in 594; then or later he received a special commission to appease the strife and write down the law. Nothing is known about the political steps leading to his appointment; perhaps the normal rivalry between powerful families grew more acute, with the risk that one group of families might exploit the grievances of the poor. The nature of the economic predicament and of Solon's solution appears in part in one of his poems. He says that he removed the "marking stones" (*horoi*), which had held the land in slavery; he adds that he freed men who had formerly trembled at the whims of their masters. These freed men should be identified with the *hektēmoroi,* or "sixth-parters," who according to Athenian tradition tilled the soil for wealthy landowners before Solon's reforms and paid a sixth of the produce to the landowners.

Solon's solution to economic problems

Thus a tolerably clear picture emerges, but the origin of the dependent status of the *hektēmoroi* is another question. Athenian scholars of the 4th century BC held that this status arose from debt, and many modern theorists have made the same assumption. One of the more sophisticated of these theories says that, as population grew, the land was cultivated more and more intensively. Eventually the peasant found that his crop would not suffice for his family until the next harvest; he therefore borrowed from a wealthy neighbour and expected to reap a surplus in the coming year. But because the need to fertilize the soil or to let it lie fallow was not understood, the next year's crop was inadequate; soon the peasant contracted a permanent tie with his wealthy neighbour, paying him a sixth of the crop, and a marking stone was placed on the land to show that it was now under an obligation.

Solon, however, does not mention debt in the extant poems, and perhaps an alternative explanation should be preferred. This says that the position of the *hektēmoroi* was not a recent development due to economic pressures and debt but a traditional status of inferiority, which arose in the "dark ages," when humble men offered their services to powerful men in return for protection in unstable conditions. A dependent status of this kind is attested in several Greek states, including Argos, Sicyon, and Thessaly. Solon in any case abolished the status of the *hektēmoroi,* making them fully free citizens and probably granting them full ownership of the land from which they had previously paid part of the crop. The wealthy landowners thus lost their claim on the *hektēmoroi* and gained little beyond freedom from the threat of violent revolution. The reform was called the *seisachtheia,* or "throwing off of burdens."

Solon tried to draw up a comprehensive statement of the law, and his laws were written on wooden *axones* and *kyrbeis,* objects of disputed character. Some of his provisions are quoted by the Attic orators and by later writers. Athenians in Classical times, however, tended to attribute all their laws to Solon, and thus the question of authenticity is sometimes difficult.

Penalties in Solonian law

The recent study of indisputably genuine provisions has shown that the penalties known to Solonian law were negative. The state could withdraw legal protection from a culprit, thereby making him an outlaw (*atimos*), or it

could utter a curse against him; correspondingly, fines could be imposed only in an indirect way—that is, the culprit could buy himself free from the basic penalties of outlawry or the curse. Positive penalties such as death, imprisonment, and confiscation of property arose only later in Attica, when the state learned to intervene actively. Few things illustrate more clearly the weakness of public authority in the early 6th century.

Some of the laws may have been intended to improve the economy. One of them offered citizenship to those aliens who settled in Attica with their families in order to practice a craft. Another, cited from the first of Solon's *axones,* allowed export of olive oil but forbade export of any other agricultural produce. The effect of such measures should not be exaggerated. They may have been traditional rules, which Solon reduced to writing. Even if they were innovations, it would be naïve to attribute the growth of Athenian prosperity primarily to them; the *seisachtheia* may have increased freedom of movement, but the forces determining the growth of the 6th-century economy were more complex than contemporary legislation. The tradition that Solon changed the standard of Athenian coinage—from that current in much of the Peloponnese to that of Corinth and Euboea—should be rejected; Athenian coinage did not begin so early.

Division into property classes

Modern writers have often attached large significance to those Solonian measures that affected the constitution. The most important of these was the division of the Athenians into four property classes. The first class, the *pentakosiomedimnoi,* consisted of those whose estates produced at least 500 measures of grain, wine, and olive oil annually. Membership in the next two classes, the *hippeis* and the *zeugitai,* depended on annual production of 300 and 200 measures respectively. Those who lacked sufficient land to produce 200 measures annually constituted the fourth class, the *thētes.* Political privilege was graded according to the classes. Only *pentakosiomedimnoi* qualified for the office of treasurer of Athena; the nine archons were drawn probably from the first two classes; the *thētes* enjoyed no privilege beyond membership in the assembly.

A widely held theory maintains that the property classes were a major innovation. It is supposed that after the unification of Attica was completed, political privilege was restricted to a hereditary ruling class, or caste, the eupatrids; they alone, it is said, were admissible to the nine archonships and hence to membership on the council. By the early 6th century some families outside this caste had achieved comparable or even greater wealth and demanded a share in political power. Solon conceded this demand, replacing heredity with wealth as the qualification for office. By so doing he transformed Athens from a closed to an open society.

Several criticisms may be made of this theory. Some of the evidence for the existence of a eupatrid class is poor, and the better evidence does not show when that class monopolized political office; it is conceivable that the eupatrids ruled Athens only at a very early stage, when the long process of unification had scarcely begun. Moreover, as Athens gradually absorbed the rest of Attica, it must have made compromises with the groups ruling at least some of the other communities; the political institutions of Athens and the other towns must have been adjusted mutually to some extent. This process would be likely to break down the exclusiveness of a hereditary ruling class in Athens; and compromises between the leading families of different towns may have even led to the adoption of wealth as the criterion for political privilege. Indeed, study of the names of the property classes suggests a pre-Solonian origin. The name of the first class, *pentakosiomedimnoi,* refers etymologically to qualification by 500 measures, but the names of the other classes are not derived from words for numbers of measures: *hippeis* refers to cavalry service; *zeugitai,* to a yoke of animals. Because the names are thus disparate, Athenian law probably accepted them at different times; in other words, at least the three lower classes were probably recognized as the basis for graduated political privilege before the work of Solon.

The poem of Solon that mentions marking stones appears complete and provides a summary of his work. It devotes 17 lines to the *seisachtheia* but only two lines and a word to his other legislative work; in those two lines he says merely that he "wrote laws alike for the bad man and the good, fitting straight justice into each." If he had carried out a major innovation by introducing the property classes, surely he would have had something to say about it.

Authenticity of the Council of Four Hundred

The same argument may be used against the attribution to Solon of another major innovation—the Council of Four Hundred. Aristotle says that Solon created such a council, drawing 100 members from each of the four tribes into which the Athenians were traditionally divided. Plutarch adds that the task of this council was to prepare business for the assembly. But neither Aristotle nor Plutarch explains how the members of the council were selected or how long they stayed in office; thus neither has any clear conception of the council. There is no certain reference to the activity of this council in the 6th century; and that is strange if Solon really created such a council, for so serious a step ought to have had some perceptible effect. The question of the authenticity of the Council of Four Hundred attributed to Solon has not been finally resolved; perhaps the easiest hypothesis is that the story of that council is a later fiction, designed perhaps in the 4th century to provide justificatory precedent for the Council of Five Hundred, which Cleisthenes introduced.

As comparisons with other ancient states at a similar stage suggest, Solon may have conceived his task for the most part as one of writing the laws down, not of making new laws. Even in the *seisachtheia* he may have believed that he was restoring a pristine condition. Reducing customary law to writing is a major improvement because it diminishes the liberty of judges to favour their friends. It also brings new precision into the law. For example, lawsuits, in the time of Solon, except for cases of homicide, were heard by the nine archons separately (each of the nine archons decided one group of cases), but serious cases could then be referred to the assembly, which was called the *hēliaia* when it sat as a court. In writing down the laws, Solon had to specify how serious an issue had to be before it could be referred to the *hēliaia.* A written statement of the law was a more valuable contribution to the growth of the Athenian state than were the deliberate innovations sometimes attributed to Solon.

Peisistratus and his sons. Thanks to Herodotus, Athenian political conflicts become clear toward 560 BC. The main rivalry was between a group called "the men of the plain," led by Lycurgus, and one called "the men of the shore," led by Megacles. But Peisistratus, who had won popularity by commanding a successful campaign against Megara, organized a third group—"the men from beyond the hills." Pretending to have been wounded by his enemies, Peisistratus persuaded the assembly to grant him a bodyguard of club bearers; with these he seized the Acropolis in 560. But within a few months, Megacles and Lycurgus joined forces and drove out Peisistratus, who withdrew, probably not from Attica but to his private estates at Brauron.

The rise of Peisistratus

A few years later Megacles quarrelled with Lycurgus and allied himself with Peisistratus, who agreed to marry Megacles' daughter. The confederates dressed up a woman as Athena, and she conducted Peisistratus back to Athens. But in 556 Peisistratus quarrelled with Megacles, who withdrew his support; and so this second tyranny collapsed. This time Peisistratus left Attica and stayed away for 10 years; he spent much of the time at Mt. Pangaeum in Thrace, where he exploited the silver mines. Several powers worked for his restoration. Thebes, among other cities, gave him money, Argos sent a force of troops, and Eretria let him use its territory as a base. In 546 he sailed from Eretria and landed above Athens, at Marathon, unopposed. When he began marching toward Athens his enemies brought their forces from the city against him; but he defeated them at Pallene and met no further resistance. The tyranny that Peisistratus thus won was passed on to his sons after his death.

Attempts have often been made to explain the conflict between Peisistratus and his rivals in terms of constitutional programs, class interests, and even differences of

The conflict between Peisistratus and his rivals

foreign policy; but such theories are not convincing. The readiness with which Megacles, Lycurgus, and Peisistratus could make and break alliances suggests that the followers of each were bound firmly to the person or family of their leaders rather than to any abstract principle, program, or interest. Moreover, any explanation of the rise of Peisistratus must do justice to the regional names that Herodotus records for the three parties. In Attic usage, the word plain meant the plain of the city of Athens; evidently Lycurgus' men belonged there. The "shore" was a long coastal strip, beginning perhaps within the city plain and extending southeast toward Sunium. Megacles came of the Alcmaeonid family, which belonged in the city during the 5th century; his party of the shore probably had its centre in or near the plain of the city. The party of Peisistratus is called by Herodotus "the men from beyond the hills." (Aristotle calls it "the men of the hills," but this term is less likely to be accurate because by Aristotle's time the tradition had become contaminated by speculation.) Brauron, the home of Peisistratus, was indeed separated by hills from the city of Athens, as was Marathon, where Peisistratus landed unopposed in 546.

The rise of Peisistratus is a clear instance of "regionalist" conflict. Previously there had been moderate rivalry between the parties of Lycurgus and Megacles, both based in or near the city, for the unification had brought ascendancy to men of that district. Bringing his local following to Athens, Peisistratus injected a new acerbity into the political struggle; he expressed the resentment toward leading families of the city felt by leading families of eastern Attica. To say this is not to deny that he won popularity outside his native district; that is proved by the grant of a bodyguard in 560, and after the Battle of Pallene he was accepted by the great majority of Athenians.

Rule of the Peisistratids

Peisistratus ruled until his death in 528 and was succeeded by his eldest son Hippias, who allowed some influence to his brothers. Later Athenian tradition was to speak favourably of the tyrants; Herodotus, Thucydides, and Aristotle said that they ruled well and observed the laws. The tyrants ensured, however, that each year some of their supporters should be among the men elected to office; hence, through the nine archonships there arose within the council of the Areopagus a group of men fully committed to the tyrants, and this may have been the key to their continued rule. Peisistratus and his sons tried to conciliate as many Athenians as possible. Herodotus says that the Alcmaeonids and some others withdrew into exile in 546, and this is credible. But the Alcmaeonids were persuaded to return before 525, when their leader, Cleisthenes, was archon.

It is often asserted that on winning his third tyranny, Peisistratus seized land from his leading opponents and distributed it to peasants. This is a modern conjecture unsupported by ancient evidence, although it is not unlikely; a possible redistribution of land was much discussed in 6th-century Athens. Peisistratus introduced the first direct taxation in Attica; he exacted a proportion of produce, recorded variously as one-twentieth (by Thucydides) and one-tenth (by Aristotle). Aristotle says that Peisistratus created judges to tour the countryside (*dikastai kata dēmous*), so that the state provided jurisdiction even in distant parts of its territory. The same office was certainly created in 453, and it is difficult to see how it could have lapsed before then. Thus the attribution to Peisistratus may be apocryphal; but if it is genuine, the establishment of this office in the 6th century was an important step toward increasing the unity and internal power of the state.

An ambitious foreign policy gave the Athenians an increased chance to take pride in their city. The same can be said of the building policy of the tyrants; although little can be asserted about this with confidence, at least a temple was built on the north side of the Acropolis to house the ancient wooden image of Athena. Magnificence, a common note of tyranny, is apparent among the Peisistratids; such poets as Anacreon of Teos, Lasus of Hermione, and Simonides of Ceos were invited to sojourn at their court.

Fall of the tyranny

An incident occurred in 514 that began a series of events that brought about the fall of the tyranny. Two men—Harmodius and Aristogeiton—felt they had been slighted by Hipparchus, a younger brother of Hippias, and they plotted to overthrow the family. But their plans were ill conceived; they succeeded only in killing Hipparchus, and in the consequent investigation Hippias came to believe that discontent was widespread. He therefore made his rule more severe, killing many Athenians. The Alcmaeonids withdrew into exile again. A year or so later they led a party of exiles into Attica and fortified a position at Leipsydrium, but they were driven away by the forces of Hippias; evidently the bulk of the Athenians was not willing to turn against the tyranny.

The Alcmaeonids resorted to diplomacy. They won the favour of the Delphic oracle, which accordingly urged the Spartans repeatedly to overthrow Hippias. A first Lacedaemonian expedition was sent by sea under Anchimolius, but it was unsuccessful; Hippias had gained cavalry from his allies, the Thessalians, and when Anchimolius landed his force in the bay of Phaleron, the cavalry overcame him. In 510 the Spartans sent a land expedition under King Cleomenes I. He defeated the Thessalian cavalry and besieged the Peisistratids on the Acropolis until they yielded and agreed to leave Attica; they withdrew to Sigeum.

The above narrative of the overthrow of Hippias is the Athenian account, the only one to survive. It emphasizes the role of the Alcmaeonids, but it does not conceal the fact that the fall of the Peisistratids was due to foreign intervention; the step fatal to Peisistratid rule was a decision taken in Sparta, although the Delphic oracle was useful in providing religious justification.

The reforms of Cleisthenes. Some time after the fall of Hippias, rivalry developed between Cleisthenes and Isagoras, who was archon in 508. Cleisthenes was worsted, but he instituted reforms that gave him ascendancy. So Isagoras appealed to Cleomenes, with whom he had contracted friendship during the campaign that overthrew Hippias. Cleomenes sent a herald and then came himself to Athens; he demanded the expulsion of Cleisthenes and of 700 families descended from those who had become accursed by slaying the followers of Cylon (see above *Athens: Early period*). Further, he tried to dissolve the council and replace it with 300 supporters of Isagoras; but the Athenians rallied to the support of the council and expelled the Spartan king. Cleomenes tried once more to make Isagoras master of Athens: he brought the army of the Peloponnesian League to Eleusis and persuaded the Boeotians and Chalcidians to attack Attica from the north and the northeast, respectively. But at Eleusis the Corinthian contingent in the Peloponnesian force mutinied, and the force disbanded; then the Athenians defeated the Boeotians and Chalcidians successively.

Rivalry between Cleisthenes and Isagoras

It is difficult to ascertain with clarity the nature of the rivalry between Cleisthenes and Isagoras. Inferences from the measures of Cleisthenes are not easy to draw, because the program of Isagoras is not known. But Isagoras' local origin may provide a clue. Herodotus observed that the relatives of Isagoras worshipped Zeus Karios. In Attica the cult of Zeus Karios is attested only at the village of Icaria just north of Mt. Pentelicus. It is therefore likely that Isagoras came from an easterly district separated by hills from the city of Athens. The possibility must be considered that his conflict with Cleisthenes was "regionalist" in character. Perhaps Isagoras attracted the support of those Areopagites who had become councillors through their loyalty to the Peisistratids; not many men had been expelled with Hippias in 510.

New tribal system

The reforms of Cleisthenes introduced a new tribal system. Previously Athenian citizens had formed four hereditary tribes, but thenceforth these were retained only for some religious purposes. Cleisthenes introduced a division of the citizens into ten new tribes, constructed in a complex way. The basic unit was the deme—a village or parish; all Attica, including Athens itself, was now divided into more than a hundred demes, which varied in size. To form tribes from the demes, Cleisthenes divided Attica into three large districts, called "the city," "the coast," and "the interior." Each of these districts was subdivided into ten *trittyes,* or "thirdings"; their boundaries were drawn so that each *trittys* held one or more complete demes. Then

to form each tribe Cleisthenes put together three *trittyes,* drawing one from each of the three districts.

Each deme had its own assembly and its annual headman, or *dēmarchos.* It kept a list of its adult male members, and membership in a deme constituted proof of citizenship. In the first instance men were assigned to demes because of local residence; but for the future, membership was hereditary. The *trittyes* failed to develop any extensive functions and served essentially as a means of distributing demes among tribes. Each tribe had its own officers and assembly. Moreover, the tribes were used for distributing the burden of military service; thenceforth the main Athenian land force was ten regiments of hoplites.

Other ancient states underwent a change from hereditary tribes to territorial tribes when the former became seriously unequal in size and could not absorb new immigrants; the introduction of territorial tribes served to equalize burdens. Considerations of this kind may have influenced Cleisthenes, but they do not explain the peculiar tripartite nature of the new Attic tribes. Some historians, unable to discover any precise purpose, have followed Aristotle in saying that Cleisthenes sought merely to mingle together citizens of different districts and thus break down local barriers. But this explanation is inadequate; the reforms gave Cleisthenes ascendancy over Isagoras, and so they must have had a partisan character. Some recent historians have observed that in outlying parts of Attica the boundaries of some *trittyes* were drawn in such a way as to break down powerful units. For example, in the northeast a unit of four townships—called the "Tetrapolis" and including Marathon—continued to exercise religious functions, but it was split between two *trittyes* and hence between two tribes. Such steps were doubtless deliberate, but they could have been achieved through unitary tribes; they do not explain why each tribe had demes from different districts.

Explanations for tribal systems

No fully satisfactory explanation for the Cleisthenic tribal system has been offered. Yet it must be admitted that the tripartite composition of the tribes looks like a measure taken in a "regionalist" struggle; it may have promoted the ascendancy of the leading families residing in or near the city. How it might serve this end can only be discerned in part. With one exception the assemblies of the several tribes met in the city; thus influential families of the city could bring their dependents to vote more easily and in greater numbers than could influential families of outlying districts. Again the rule of candidature for office may be significant. In 501 the office of general (*stratēgos*) was introduced or reformed; thenceforth each year ten generals were elected by the assembly of all citizens, but one had to be chosen from each tribe. Likewise the number of archons was increased from nine to ten through the addition of the secretary of the *thesmothetai,* although they continued to be called the nine archons. If one member of the board of ten generals was to be chosen from each tribe, then, because each tribe held some city territory, all members of the board might be city men; and this was likely to happen because the electoral assembly met in the city.

Council of Five Hundred

The other major reform of Cleisthenes was the introduction of the Council of Five Hundred. The members of this were chosen by lot and held office for one year; in the 4th century one second term was permitted, but at first even this may not have been allowed. Each tribe supplied 50 members to the council, and within the tribes each deme supplied a number of councillors roughly proportionate to its size. The task of the council was called *probouleusis* and consisted of preparing business for the assembly; the rule was that no item could come before the assembly unless it had been considered previously by the council.

The reason for establishing the new council is not difficult to discern. As Aristotle observed, there had to be a smaller council to prepare items for a public assembly. A mass meeting of several thousand could not deal with business in the raw. But the preparatory council, if it had the will, was in a position to influence the outcome of the assembly's vote by its formulation of the issues. The members of the Council of Five Hundred, however, were not men of particular talent or ambition because they were chosen by lot; and they did not acquire expertise because they served only for one year or, at most, two. Thus by virtue of its personnel the council was not likely to assert itself. In practice it prepared business efficiently but left the real decision to the assembly; its existence ensured that no other body would encroach on the assembly's authority.

The practice of ostracism

Aristotle, perhaps correctly, attributed to Cleisthenes a further measure—the introduction of ostracism. By this practice, in late winter of each year the assembly was asked whether or not it would hold an ostracism; if it answered affirmatively, at the next possible meeting each citizen deposited a potsherd on which he had written the name of the man who he would most wish to have leave Attica. A valid ostracism required 6,000 votes. The man against whom the most votes were cast had to leave Attica for ten years but retained possession of his property. Attempts to link the presumed reasons for this custom with the other measures of Cleisthenes are not persuasive. The origin and purposes of ostracism remain opaque.

In the second half of the 5th century, Cleisthenes was regarded as the founder of the democracy. The introduction of the new council may have been even more effective than the tribal reform, both for the outcome of the struggle with Isagoras and for the further development of Athens. The council ensured that real decisions were taken in the assembly, and in a sense this was democratic. But the word *dēmokratia* had not yet been coined in the time of Cleisthenes, and there was no synonym; his purposes should not be sought in democratic ideas. The assembly met in the city, and so it could most readily be influenced by families of the city and its immediate neighbourhood. Cleisthenes brought "regionalist" struggles to an end by winning victory for the city.

Society and the economy. In antiquity the Athenian economy depended upon the natural resources of Attica. Noteworthy mineral resources included silver in the southeast and clay for pottery. Attica's pronouncedly Mediterranean-type climate—mild temperatures and less rainfall than in many parts of Greece (the annual average is 16 inches, but there is considerable variation from year to year, and virtually all the rain comes in winter)—suited the territory to fruits such as the vine, the olive, and the fig, which can send roots deep down into the soil and in summer draw on moisture stored up from the previous winter.

Economy in the late 7th century

In the late 7th century BC, Attica was predominantly a land of subsistence farming. The main crop was grain—especially barley, which is hardier than wheat—but vines were also grown. There was some fishing. Relatively few animals were kept. Marginal land was brought under cultivation to support the growing population. But as the record of black-figure ware shows, exports were beginning to expand even before 600. Furthermore, there was contact with the flourishing commercial cities of Corinth, Megara, and Aegina. The crisis facing Solon was due not only to the poverty that arose when population grew beyond the local means of subsistence but also to the desire to achieve better conditions.

The measures taken by Solon are now obscure, but something can be said about their effects. Thenceforth rural Attica was a land of free peasants; there is no later trace of any quasi-servile status, such as that of the *hektēmoroi.* On the other hand, wealthy men could still command a large following of humble dependents. (As late as 480, at the Battle of Artemisium, Cleinias supplied his own trireme, which he equipped and manned at his own expense; this means that he could command about 200 able-bodied men.) By abolishing the status of the *hektēmoroi,* Solon increased the mobility of rural manpower; thenceforth a subsistence farmer who failed might drift to the city. A tradition about Peisistratus may indicate some growth of an urban proletariate; Aristotle says that the tyrant lent money to poor men so that they could settle as farmers. Such a scheme probably concerned newly cleared land on the hills, where vines and olives could be grown but required capital.

Growth of trade

The distribution of black-figure ware is the clearest indication of the growth of the Athenian economy. In the period 560–520, Attic vases reached Chios, Lesbos, Cyprus, Asia Minor, the coasts of the Black Sea, Egypt, south Italy, Etruria, and southern France. (Many of the

vases were of high artistic quality and prove the existence of skilled potters in Athens.)

Some political developments helped the trade. Sometime between 600 and 560 the Athenians captured the island of Salamis from Megara. Thus, the bay of Phaleron became safe from attack, and trade with Corinth was facilitated. Athenian goods reaching the west may have travelled in Corinthian ships. After Peisistratus had finally made himself tyrant, an Athenian expedition under Miltiades settled the Thracian Chersonese on the European shore of the Hellespont, and the tyrants kept up the connection with the settlement into the next generation. Peisistratus captured Sigeum in the Troad from the Mytilenaeans and sent one of his sons to rule there as tyrant. The expeditions to Sigeum and the Chersonese suggest that Peisistratus was interested in the trade route through the straits to the shores of the Black Sea.

The effect of coinage on the Athenian economy is difficult to determine. Studies conducted since World War II have shown that the earliest coins were struck in Lydia *c.* 640–*c.* 630, not *c.* 700 as previously supposed. The invention was adopted gradually by many Greek cities. The earliest Attic coins, called by numismatists "Wappenmünzen," bore a variety of heraldic designs, which may stem from different families; these were struck not before *c.* 575. They were superseded by coins bearing the design of the owl, the emblem characteristic of Athens, and these began not before *c.* 527. Some scholars indeed would put the earliest "Wappenmünzen" soon after 550 and the earliest "owls" *c.* 510.

While in exile, Peisistratus drew silver and perhaps gold from the mines of Mt. Pangaeum; this enabled him to raise forces for his restoration and to maintain a bodyguard thereafter. He may have begun working the silver mines at Laurium in southeast Attica. The Peisistratids were not the only family to interest themselves in foreign sources of wealth. A legend of the Alcmaeonids told how one of their ancestors, Alcmaeon, visited the Lydian king Croesus and came away laden with gold. Part of a grave monument honouring an Athenian named Croesus and erected *c.* 530 has been found to the southeast of Mt. Hymettus and has been claimed as indicating the home of the Alcmaeonids. This inference is not necessary; there is no reason to suppose that the Alcmaeonids were the only Athenian family to develop relations with the Lydian kingdom.

Foreign sources of wealth

Although the stages of economic growth are obscure, by 500 Attica was a very different place than it was a century before. It had become the leading exporter of vases among the Greek cities; its other exports included oil, wine, and silver. It imported grain, fish, and ship timber. The population was growing, and economic expansion had attracted immigrants. According to Herodotus, toward 500 the number of adult male citizens was 30,000; the reorganization by Cleisthenes had entailed some survey of the citizen body, and this figure may be approximately correct.

Other important poleis. The Greek *polis* was usually small, and therefore special interest attaches to those *poleis* that achieved rule over a comparatively large territory. Chief among these were Sparta, Athens, and Thebes. Lacedaemon was, strictly speaking, a federal state, although most power belonged to the dominant city, Sparta. Athens organized Attica as a unitary state. Boeotia, however, in which Thebes was located, was a federation of about a dozen cities that varied greatly in size and power.

Thebes. Geographically, most of Boeotia consisted of two river basins—those of the Asopus in the south, where Thebes predominated, and of the Cephissus in the north, where Orchomenus was the strongest city. There was rivalry between Thebes and Orchomenus, and usually Thebes had the superiority. Starting in or before the 6th century BC, the Thebans built up a federation of Boeotia, which was dissolved in 479. Nothing is recorded about its constitution, but it is fair to infer from the structure of a restored confederacy (447–386) that it was organized on relatively liberal principles. (The constitution of the restored Boeotian confederacy is known and has remarkable features: each city managed local affairs, and the supreme federal organ was a representative council of 660 members; in 395 Thebes supplied 240 of these—more than any other city but less than a majority.) Apparently by 519 the early federation embraced much of Boeotia. Orchomenus, however, did not belong, and on the southern border Plataea refused to join; attacked by the Thebans in 519 (or possibly *c.* 509), the Plataeans gained the alliance of Athens and hence preserved their independence.

Boeotian confederacy

Phocis and Thessaly. Northwest of Boeotia, Phocis lay as a broad wedge separating Locris into two parts. As late as the early 5th century, the men of East and West Locris were still aware of their mutual bond and cooperated to found Naupactus on the north shore of the Corinthian Gulf. For a long time in the 6th century, Phocis was on the defensive in an intermittent conflict with Thessaly; finally about 500 the Thessalians invaded Phocis, but the Phocians inflicted heavy defeats on the Thessalian infantry and cavalry in two successive battles.

Thessaly itself, a large plain surrounded by four ranges of mountains, stood apart from the political development of the rest of Greece. It was dominated by a landowning aristocracy; indeed its several cities may have been brought into existence by their local dynasties, such as the Aleuads of Larissa and the Scopads of Crannon. These cities were strongholds or market towns rather than centres of any lively political development. Intermittently, the leader of a local dynasty gained recognition of his ascendancy in all Thessaly and took the title of *tagos;* but at other times federal organization was dormant.

Overseas settlements. The Aegean islands and the cities of western Asia Minor were the scene of a particularly brilliant development in the Archaic period. The affairs of Mytilene, the chief town of Lesbos, are known in part from the fragmentary poems of Alcaeus (*c.* 620–*c.* 580). His lifetime saw the final overthrow of the Penthilids, a clan that had once controlled the city. A complex series of factional struggles within the aristocracy followed; there is no sufficient warrant for the common assumption that Pittacus, one of the contenders, led a popular movement. Alcaeus was at first an ally but later an enemy of Pittacus and was driven into exile, probably more than once. Eventually the Mytilenaeans chose Pittacus as *aisymnētēs,* a position which Aristotle described as an elective tyranny. Pittacus' primary task was to preserve the city from the threat posed by the exiles. He seems to have succeeded in bringing civil strife to an end, and he then laid down his office. Hellenistic scholars said that Pittacus ruled for ten years (590–580), but the chronology of Mytilene is difficult. Alcaeus in exile said that he longed for the assembly (*agora*) and the council (*bolla*).

Developments in the Aegean islands

Farther south, on the island of Chios, an inscription of *c.* 575–*c.* 550 attests a *bole demosie,* a term often taken to mean a "popular council" as distinct from an "aristocratic council," although other interpretations are possible; for example, the term may mean a "state council" as distinct from "local councils." At least the occurrence of the epithet suggests that there was more than one council in Chios, and this indicates some sophistication of political development.

On the west coast of the Asiatic mainland, the Ionian settlements were among the most prosperous Greek cities of the Archaic period. They held the alluvial plains at the mouths of major rivers, which flow westward into the sea, and could thus expand considerably before feeling the pressure of population on local means of subsistence. Moreover, the rise of the Mermnad dynasty in Lydia meant that the Ionian settlements had a strong, stable, and wealthy power as their neighbour; excavations at Sardis show that Lydian trade with Greek cities increased steadily during the Mermnad period. It is not surprising that the 6th-century Ionians were ahead of other Greeks in some spheres; for example, Thales of Miletus, who predicted the year of the solar eclipse of 585, began the series of Greek scientists. Miletus, the richest Greek city of the Asiatic mainland, had a tyrant named Thrasybulus, who exchanged friendly messages with Periander. Afterward the city suffered a long period of internal strife, but sometime before 500 it achieved stability by inviting mediators from the island of Paros; by the settlement these mediators made, power appears to have been entrusted to men of landed wealth.

Ionian settlements of Asia

Phocaea, the northernmost city of mainland Ionia, did much to open the western part of the Mediterranean to Greek enterprise in the 6th century. The Phocaeans developed commercial relations with the flourishing kingdom of Tartessus in Spain (*c.* 620–*c.* 540). On the south coast of France the Phocaeans founded the colony of Massilia (the modern Marseille) *c.* 600; the Massiliotes in turn founded dependent colonies as trading stations in a westward line reaching into the northeast coast of Spain. The Phocaeans also founded a settlement in Corsica called Alalia *c.* 560. When Persian forces attacked Phocaea *c.* 540, the Phocaeans withdrew by sea and, although some returned, a large party sailed to Corsica and joined the colony of Alalia. But after practicing piracy for five years, the people of Alalia were attacked by the combined fleets of the Etruscans and the Carthaginians; although the Greeks claimed to have won the battle, they had to withdraw from Corsica, and, after a sojourn at Rhegium, they settled at Elea (Velia) in southern Italy. The Etruscans took Corsica, and the Carthaginians took Sardinia, which they subdued after a struggle with the natives; together the two powers exercised ascendancy in the western basin of the Mediterranean.

Greeks in Sicily and southern Italy

In Sicily the 6th century marked the beginning of a protracted conflict between Greeks and Phoenicians. The latter held three bases—at Motya, Panormus, and Solus—in the western part of the island; they were often joined by the people of Segesta in their opposition to the Greek advance from the east. The first known conflict occurred *c.* 580, when Pentathlus of Cnidus led a force from Cnidus and Rhodes to settle at Lilybaeum—a position commanding the harbour of Motya; he was driven out by the Phoenicians and Segestans. In the next decades Carthage gained leadership over the Phoenician settlements in Sicily. In 510 Dorieus, the half-brother of Cleomenes of Sparta, led a party of perhaps 1,000 men to settle at the foot of Mt. Eryx, but he was overcome by the combined forces of Carthage, Segesta, and the Phoenicians of Sicily.

In southern Italy Sybaris, an Achaean colony, built up a loose federation of Achaean colonies from *c.* 550. Croton, another Achaean city, became its chief rival. Soon after 530 Pythagoras the philosopher settled in Croton; his political teaching is not clear, but his followers gained predominant influence in the city and set about expanding its rule. In 510 the Crotoniates took advantage of an internal dispute in Sybaris to attack and destroy the latter; at the time of its destruction Sybaris was perhaps the most thriving of Greek colonies. Croton then became the leading power in southern Italy, although the alliance it created was not so extensive as that formerly led by Sybaris. The Italiote cities played a prominent part in Greek thought and culture in the late 6th century; they produced such men as the philosopher Parmenides of Elea and the poet Ibycus of Rhegium. Croton had a flourishing school of medicine; and one of its doctors, Democedes, served for a time at the court of Darius I of Persia.

Relations among the Greek poleis to the end of the 6th century BC. *The beginning of the Archaic period.* Early in the Archaic period the structure of the *polis* was weak and therefore there was little official intercourse between *poleis.* Thucydides remarked that in early Greece there were hardly any wars sufficiently extensive to bring in many cities; he noted as an exception the war between Chalcis and Eretria, neighbouring cities on the island of Euboea, in which many other Greeks supported one side or the other. This struggle is called the Lelantine War because it was in part a conflict for control of the Lelantine Plain in Euboea. The allies of Eretria included Miletus, and those of Chalcis included Samos. Chalcis also gained help from Pharsalus in Thessaly. (Some of the issues were local; Samos and Miletus went to war for control of Priene as late as 441.) The date, duration, character, and outcome of the Lelantine War are far from clear; a good case can be made for a date in the late 8th or early 7th century BC, although a later date can also be reasonably defended. If the earlier date is correct, rivalry for colonial sites probably exacerbated the struggle.

Local leagues

Although dealings between cities were slight, local leagues arose at an early date. These were predominantly religious, but they sometimes acquired a political character. An important example was the League of Amphictyones, comprising several powers of central and northern Greece from Boeotia to Thessaly. It arose to protect the cult of Demeter at Anthela near Thermopylae, but it acquired control over the sanctuary of Apollo at Delphi. The oath of its members included a promise not to destroy or cut off the water supply of any member-city, even during warfare.

Another local league included 12 Ionian cities—Samos, Chios, and 10 cities on the Asiatic mainland. Members built a temple—the Panionion—on the northern slope of Mt. Mycale and gathered there for periodic festivals. When the Ionians felt themselves threatened by the Persian advance *c.* 544, they gathered at the temple for political deliberation.

The early tyrants in the Peloponnese brought greater cohesion to their cities and asserted themselves abroad. The colonies founded by the Cypselids and the dealings of Periander with Procles of Epidaurus have been noted above. Periander maintained good relations with Oriental powers. For example, after seizing 300 Corcyrean boys as hostages during a dispute with the island, he sent them as a present to King Alyattes of Lydia (the present did not arrive; the convoy put in at Samos and the Samians enabled the hostages to escape). Periander's nephew was called Psammetichus—a non-Greek name (Psamtik) borne by some pharaohs of the Saite dynasty (664–525), which suggests Corinthian dealings with Egypt.

First Sacred War

Cleisthenes of Sicyon fought a war against the Argives, but his great opportunity was provided by the First Sacred War. This war (*c.* 595–*c.* 586) concerned the Delphic sanctuary and the Phocian town of Crisa. The latter had a good harbour on the northern shore of the Corinthian Gulf; it controlled a fertile plain, through which ran the route taken by pilgrims to Delphi. According to the tradition—known from texts written more than two centuries later—the Crisaeans levied tolls on these pilgrims; and the Amphictyonic League, exhorted by a response of the oracle, went to war against the Crisaeans because of their impiety. The war developed into a siege of Crisa; the northern powers blockaded it on the landward side, but it received supplies by sea. Thereupon Cleisthenes joined in the war; he built a fleet and hence stopped the transport of supplies to Crisa. The town fell in 591–590, although pockets of resistance still held out. Cleisthenes received a share of the spoils and won a chariot victory at the Pythian festival of 582. The war brought him distinction in Greece, and his fleet brought him power.

The war led to reorganization of Delphic affairs. The previous condition is not clear. Thenceforth the Amphictyonic League had ultimate control over the sanctuary; the town of Delphi managed its own affairs and supplied the priestess and interpreters for the oracle. The Pythian festival was reorganized on a larger scale, with athletic and other competitions; it was to be held every four years, and the year 582 was reckoned as the first Pythiad.

Reorganization of festivals

The Pythian reorganization seems to have prompted the reorganization or foundation of other panhellenic festivals. About 580 the Eleans finally gained ascendancy over the Pisatans and reorganized the Olympic festival, held likewise every four years; there had been a festival at Olympia before, but its early history is obscure. (In the second half of the 5th century, Hippias of Elis published a list of Olympic victors; he assigned the first celebration of the festival to 776 and all subsequent scholarship followed him, but the reliability of the entries in the early part of his list is an unsolved problem.) The reorganization of *c.* 580 may have caused the festival to attract visitors from a wider area than before. In Corinthian territory the Isthmian Games were founded *c.* 580 and celebrated every two years. In the territory of Cleonae the Nemean Games were founded in 573 and likewise celebrated every two years; it is not clear whether Cleonae, which lay between Sicyon and Argos, was under the influence of Cleisthenes or of the Argives at the time of the foundation.

In general it is disputed whether panhellenic festivals owed their expansion in the early 6th century to the tyrants or to other regimes. The fact emerges, however, that the expansion of the festivals reflects the same growth

of wealth and ambition as do the early tyrannies. The four festivals—Pythian, Olympic, Isthmian, and Nemean—were panhellenic in that they admitted all Greeks without regard to such distinctions as that between Dorians and Ionians. They attracted visitors and competitors of all kinds, but their contribution to promoting panhellenic consciousness should not be exaggerated; they did not acquire political significance.

The later Archaic period. In the 7th and 6th centuries BC, the overseas activities of Greeks on the European mainland increased. The fortunes of the Asiatic Greeks, and later of those on the European mainland, were affected increasingly by the non-Greek powers of Asia Minor.

Lydian influence

The kingdom of Lydia was rich in gold and silver and in an alloy of the two—electrum—which occurred there naturally. In founding the Mermnad dynasty at Lydia, Gyges (*c.* 680–648) seems to have filled a power vacuum. He made indecisive attacks on some Ionian cities; he raided Miletus and Smyrna, and he captured the citadel of Colophon. This policy was continued by his successors: the third and fourth members of the dynasty, Sadyattes and Alyattes, fought a war of 12 years against Miletus. In each of the first 11 years the Lydian forces plundered the Milesian crops before the harvest; but the sea routes to Miletus remained open, and in the 12th year Alyattes concluded a treaty of friendship and alliance with the city. His successor Croesus attacked Ephesus and then the other Greek cities of the Asiatic mainland and made them pay tribute, although Miletus retained its privileged status; regular tribute was more satisfactory to all parties than were intermittent raids. Under Croesus the Lydian kingdom reached its widest extent, stretching as far as the Halys River in the east.

In spite of the raids and the tribute, the Mermnad dynasty brought advantages to Greeks. Trade with Lydia allowed a steady growth in Ionian prosperity. Greeks seem to have travelled freely in Lydia; later tradition told of the hospitality shown by Croesus to the Athenians Solon and Alcmaeon; and although details are suspect, such stories attest the King's philhellenism. No Lydian garrisons are recorded in Greek cities. Looking westward beyond the mainland, Croesus achieved treaties of friendship with Ionians of the offshore islands. He sent gifts for dedication in Greek sancturaries at Ephesus, at Branchidae near Miletus, at Thebes, and at Delphi. He exchanged presents with Sparta.

About 544 (the precise date is disputed) Croesus went to war against Cyrus, the founder of the Persian Empire. Cyrus drove the Lydian forces back to Sardis, besieged the city, and overcame Croesus. The Greek cities of the Asiatic mainland had to consider their attitude toward the new power approaching them from the east. Before Cyrus left Sardis, the Greeks sent envoys and asked him to let them continue in the same status they had enjoyed under Croesus. Cyrus confirmed the Milesians in their previous condition, but he would make no undertaking to the others because they had refused his request that they desert Croesus while the war was in progress.

After a short stay in Sardis, Cyrus hastened to Ecbatana in Media. The Lydians promptly rose in revolt and were joined by the Greek cities of the Asiatic mainland. Cyrus sent forces that reduced Lydia anew and then captured the Greek cities piecemeal; Miletus alone continued to enjoy special treatment. Thus, by *c.* 540 the Persian Empire had acquired the Greek cities of Asia Minor. These cities became subject to the nearest satrap, or provincial governor, such as the satraps resident at Sardis and Dascylium; and the imperial administration usually tried to maintain order by supporting a local man as tyrant in each city. The cities were required to pay tribute, but generally their burdens were not severe.

Persian rule

From the mainland of Asia Minor the Persian administration might have hoped to extend its authority to the offshore islands, but the steps taken are obscure. Details are available only for Samos. This island was ruled from *c.* 535 by the tyrant Polycrates, who raised it to unprecedented power; he practiced piracy extensively and maintained the strongest fleet in the Aegean. Polycrates seems to have taken advantage of a change in naval engineering; probably during his time the trireme was invented. His contacts were not only with Greeks. He exchanged presents with Ahmose II, the pharaoh of Egypt; later (in 525) Cambyses, the son and successor of Cyrus, launched his successful expedition to conquer Egypt, and Polycrates sent 40 triremes to help Cambyses. Toward 522 Oroetes, the satrap of Sardis, conceived a design for conquering Samos; he enticed Polycrates into a conference on the mainland and crucified him. But before Oroetes could proceed further, Cambyses was overthrown, and, after a struggle for the succession, the usurper Darius made himself king. Darius had Oroetes assassinated. Meanwhile, Maeandrius, the second in command to Polycrates, ruled Samos; but shortly afterward Darius sent a special force under Otanes to attack the island. The force was guided by Syloson, the exiled brother of Polycrates. Otanes sacked Samos and then entrusted it to Syloson.

How far Persian rule extended over Greeks *c.* 514 is known from Herodotus' account of the expedition that Darius led into Scythia. Darius conducted his forces across the Bosporus and thence northward and across the Danube; meanwhile the Greek tyrants from cities under his control sailed ahead to the Danube to bridge the river, and Herodotus gives a list of them. They included men from such cities of the Ionian and Aeolian mainland as Miletus, Phocaea, and Cyme, and also tyrants from several cities on the Asiatic shore of the Hellespont and Propontis, notably from Abydos, Lampsacus, Parium, Cyzicus, and the neighbouring island of Proconnesus. The only Aegean islands represented in the list were Samos and Chios; the tyrants present from the European mainland were Miltiades, the Athenian ruler of the Thracian Chersonese, and one from Byzantium. Elsewhere Herodotus remarks that Coes brought a force from Mytilene and that Darius made him tyrant of that city as a reward for his services on the expedition.

The foreign policy of Sparta in the 6th century was determined first by the city's position in the Peloponnese and later by its attitude toward developments across the Aegean. Sparta's wars with Tegea (*c.* 580–*c.* 550) and Argos (*c.* 544) and the growth of the Peloponnesian League were noted previously (see above *Sparta: The Peloponnesian League*). In its dealings with powers outside Greece, Sparta at first followed—consciously or otherwise—the practice of tyrants like Periander. Relations were opened with Croesus toward the middle of the century. The Lacedaemonians sent to Sardis to buy gold for a proposed statue of Apollo, and Croesus gave them the gold free. Thereafter the Lacedaemonians sent a large ornamental bowl of bronze as a present to Croesus; the bowl only travelled as far as Samos, and the Lacedaemonians said that it was seized there by Samian pirates. Relations were opened likewise with Ahmose of Egypt. Shortly before the dispatch of the bowl to Croesus, Ahmose sent the Lacedaemonians an embroidered corslet of linen, decorated with gold; this too, travelled no farther than Samos.

Spartan relations with Lydia

The tradition reaching Herodotus portrayed the dispatch of the bowl as the Spartan response to a request from Croesus for a military alliance against Cyrus. Likewise it said that when Croesus was besieged in Sardis by Cyrus, he appealed to Sparta for help, and, although the Lacedaemonians were engaged in their war with Argos, they prepared to send a fleet, but that the news arrived of Sardis' capture before the fleet could sail. These stories may have gained something in the telling. The historian should beware of statements of unrealized intentions; the fact emerging from the alleged preparation of a fleet is that Sparta did not send aid to Croesus. Even so, the exchanges of presents may have had political implications. There may have been more presents than are recorded, and at the least they indicate commercial possibilities.

Herodotus tells a tantalizing story of action taken by the Ionians immediately after Cyrus overcame Croesus. Allegedly they gathered at the Panionion and sent envoys to Sparta for help. The Spartans decided to give no help, but they sent a penteconter (50-oared galley) to Phocaea and one of its crew proceeded to Sardis. There he gave Cyrus the message of the Lacedaemonians—that they would not tolerate the Persians ravaging any Greek city—and

Cyrus made a contemptuous reply. In this story it is not difficult to recognize an element of fiction. The account of the Lacedaemonian penteconter and message to Cyrus was surely invented later to gloss over the Spartan refusal of aid to the Ionians. But the Ionian appeal to Sparta may well be authentic; it attests the standing that the Lacedaemonians had achieved in Greece and the extent to which they had interested themselves in affairs beyond the Aegean.

Decline of Spartan overseas operations

Such a development of Lacedaemonian interests had been possible in the time of Croesus; accordingly it was checked by the Persian conquest of Lydia. A further check to Spartan enterprise in the eastern Aegean occurred *c.* 525. In response to the appeal of some Samian exiles, the Lacedaemonians and the Corinthians sent a fleet to try to overthrow Polycrates. The force besieged Samos for 40 days unsuccessfully; then it withdrew.

Modern historians have sometimes supposed that in consequence of the final subjugation of Messenia, Sparta concentrated on developing its power by land and declined to acquire interests overseas. This is incorrect. In the middle of the 6th century Sparta concerned itself with developments across the Aegean, but by 520 its attempts to look eastward had suffered more than one setback. Thereafter until 412, it made little attempt at naval enterprise and concentrated on extending its influence by land operations. Such a policy appears in the reign of Cleomenes I (*c.* 520–*c.* 490), but before considering his early activities, attention must be given to the foreign policy of Athens.

Late in the 7th and early in the 6th century, the Athenians scarcely developed foreign interests beyond border warfare with their immediate neighbours. There was a long struggle with Megara, the main issue being control of the island of Salamis. During this war Peisistratus led a campaign sometime before 560, in which the Athenians captured Nisaea, the eastern harbour of the Megarid; it is not clear when the Megarians recovered Nisaea. The main outcome of the fighting was that the Athenians won Salamis.

There are slight traces of Athenian enterprise farther afield. An Athenian contingent, commanded by Alcmaeon, took part in the First Sacred War; thenceforth, Athens had a vote in the Amphictyonic Council. A less reliable tradition said that the original resolution of the Amphictyones against Crisa was proposed by Solon, but this may be a later invention of Athenian pride. Many historians believe that late in the 7th century an Athenian force commanded by Phrynon fought the Mytilenaeans for control of Sigeum in the Troad. The historicity of this war depends on a complex problem of chronology: if indeed there was such fighting at so early a date, then it was probably not an official act of the Athenian state but the work of some Athenians. Certainly the question of Sigeum arose, or arose again, in the time of Peisistratus.

Athenian foreign policy may be said to begin with the final tyranny of Peisistratus and his sons (546–510). Peisistratus established an Athenian presence in the neighbourhood of the Hellespont, he asserted Athenian influence in the heart of the Aegean, and he maintained good relations with several powers of the Greek homeland. In the first of these spheres he captured Sigeum from the Mytilenaeans and entrusted it to Hegesistratus, one of his sons. Hegesistratus had to conduct further fighting against the Mytilenaeans, but he remained in control of Sigeum; indeed, the Peisistratids withdrew there when they were driven from Athens in 510.

Athenian presence in the Hellespont

The foundation of an Athenian settlement in the Chersonese on the shore of the Hellespont was still more ambitious and was probably carried out soon after 546. The settlement was invited by the natives of the Chersonese—the Thracian Dolonci—who desired reinforcements against the attacks of their neighbours, the Apsinthii. The settlement was led by Miltiades of the Philaid clan, which apparently belonged to Brauron because the deme in which Brauron stood was called Philaidae. After bringing Athenian settlers, Miltiades built a wall across the isthmus of the Chersonese and waged war indecisively against Lampsacus on the other shore of the Hellespont. At one stage in the war Miltiades was taken prisoner by the Lampsacenes, but he was released through the intercession of Croesus. At his death Miltiades entrusted the Chersonese to Stesagoras, the son of his uterine brother Cimon. The war with Lampsacus continued and in consequence of it Stesagoras was assassinated. Thereupon (*c.* 516) the sons of Peisistratus sent out from Athens another son of Cimon, who bore the name of his uncle Miltiades. The dispatch of the younger Miltiades shows that the Peisistratids were keen to preserve their connection with the Chersonese.

In the heart of the Aegean, Peisistratus had a political debt to pay. In 546 Lygdamis, an adventurer from Naxos, helped him with men and money in his final restoration. In return Peisistratus subdued Naxos and entrusted it to Lygdamis. (The latter helped Polycrates to make himself tyrant of Samos.) Peisistratus also carried out a ceremonial purification of the island of Delos, the main sanctuary of the Ionians, by digging up and removing all corpses buried within sight of the temple; by the purification Peisistratus took advantage of the Athenian claim—first attested in the poems of Solon—to be the metropolis of all other Ionian cities.

Athenian alliances

Several cities in the Greek homeland aided Peisistratus in his final restoration; they included Thebes, Argos, and Eretria. One of his sons bore the name Thessalus, which suggests that the family had ties with the Thessalians; Hippias' alliance with Thessaly was apparent in the last years of his rule, when Thessalian cavalry helped him against Lacedaemonian attacks. The Peisistratids seem to have also enjoyed the friendship of Sparta for a time; some years after the overthrow of Hippias the Spartans conceived a plan for restoring him and remembered that the Peisistratids had been their friends. (International "friendship" was a formal matter, akin to alliance, and Spartan recollection may have been correct.)

In large part the tyranny owed its foundation and its preservation for many years to the favour of other cities; neighbours of Attica, like Thebes and Eretria, and powers farther afield were keen to see a *persona grata* in control of Athens. Developments abroad eventually weakened the Peisistratids. A Spartan force overthrew Lygdamis of Naxos, perhaps during the return of the unsuccessful Spartan expedition against Polycrates of Samos, the friend of Lygdamis; the attack on Lygdamis may not have been intended to harm the Peisistratids, but it removed a friend of theirs. A few years later the Persians overthrew Polycrates.

Some time later the Plataeans were attacked by the Thebans; on the advice of Cleomenes, who happened to be near with a Lacedaemonian force, the Plataeans sought and gained the alliance of Athens. (The traditional date of concluding this alliance is 519; it has been disputed but cannot be proved wrong.) Herodotus, recounting the alliance, says that the Lacedaemonian aim was to create discord between Athens and Thebes. That may be an anachronistic inference from conditions obtaining in his own time, but the conclusion of the alliance did create discord between Athens and Thebes; the Athenians promptly sent a force that defeated the Boeotians.

If the Plataean alliance was concluded in 519, then the international position of Hippias had been impaired well before 514. After the assassination of Hipparchus, Hippias sought new friends abroad; he gave his daughter in marriage to the son of Hippoclus, the tyrant of Lampsacus. According to Thucydides' recording of the marriage, Hippias knew that the Lampsacene tyrants had influence with Darius. Hippias may also have recognized that the makers of Persian policy were showing a new interest in the West. At least Hippias secured the prospect of a refuge in the Persian Empire, should he be overthrown.

Spartan policy, as it developed in the early years of Cleomenes, followed recognizable principles. On the one hand Sparta gave up its attempts to extend its influence by naval enterprise across the Aegean. This meant that Sparta would not oppose the expansion of the Persian Empire, provided that it did not reach European Greece. Thus when the Persians under Otanes finally attacked Samos, Maeandrius sailed to Lacedaemon and appealed to Cleomenes for help; instead, Cleomenes advised the ephors to expel Maeandrius from Lacedaemon, and this was done. Again, when Darius had attacked the Scythians,

the latter sent envoys to Sparta and offered an alliance for military operations against the Persian Empire, or so Herodotus says; the account may be imprecise, and the only result was that Cleomenes learned from the envoys the habit of drinking unmixed wine.

On the other hand, the Spartans tried to enlarge their ascendancy by land operations on the Greek mainland; this led them into conflict with friends of Persia, when Persian influence appeared to encroach on European Greece. The alliance of Hippias with Hippoclus of Lampsacus—a friend of the Persians—may well have induced the Spartans to adopt extreme measures against Hippias, measures that culminated in the expulsion of the Peisistratids from Athens. Whether Athens became a member of the Peloponnesian League in 510 is perhaps a question of terminology. Certainly in the ensuing decade the Spartans tried repeatedly to ensure that there should be a favourable regime in Athens.

Spartan–Athenian antagonism

As related above (see above *Athens: The reforms of Cleisthenes*), some time after the Peisistratids' fall Cleomenes tried to support Isagoras against Cleisthenes, first by sending a herald and then by coming himself to Athens, where he attempted to dissolve the council; but the Athenians drove him out. Thus the Athenians had quarrelled with Sparta, and so they turned to the alternative source of support; they sent envoys to Sardis to seek an alliance with Persia, and, at the request of the satrap, the envoys gave earth and water, the tokens of submission. Hence for a time Athens was technically a vassal of Persia.

As also mentioned previously, Cleomenes responded to the rebuff he received at Athens by calling out the army of the Peloponnesian League and leading it against Attica, but at Eleusis the Corinthian contingent mutinied and the force disbanded. This operation caused a crisis in the league, and a congress of envoys from the member-states was summoned. Doubtless questions of authority and command were discussed; indeed, this may have been the occasion when a multilateral agreement was first accepted, superseding the bilateral treaties from which the league had arisen. For the congress the Spartans invited Hippias from Sigeum, and they proposed to the allies that he should be restored to Athens. But first the Corinthians and then the other allies rejected the proposal; it was abandoned and Hippias returned to Sigeum.

As a result of the congress, the Spartans gave up trying to control Athenian affairs. The Athenians, however, soon became perturbed because asylum was granted to Hippias within the Persian Empire, and they sent envoys to Sardis to complain. The satrap replied with a brusque instruction: they should restore Hippias, if they cared for their safety. The Athenians were not willing to comply, and thus their close relationship with the Persians came to an end.

The 6th century saw two major changes in the balance of power in European Greece. One of these was the development of the Peloponnesian League; by 500 it was the largest single force in interstate affairs, capable of providing a nucleus for resisting a foreign invader and fit to serve as a model for other alliances. The other change was the rise of Athens to mature statehood. Before 600 the Athenians had done little to assert themselves beyond the borders of Attica, and for a long time afterward Athens was much subject to foreign influence; other cities played a large part in the final restoration of Peisistratus and even in the overthrow of Hippias. But the ambitious policy of Peisistratus had aroused new aspirations among the Athenians; in the decade following 510 they came to assert themselves against Sparta and Sardis, and thus they achieved full independence. (B.R.S.)

THE 5TH CENTURY BC

The period of the Greco-Persian wars. *The Ionian revolt (499–493 BC).* The sequel to the conquest of Lydia by Cyrus the Great was the subjection of the Ionians and other east Greeks. The Persians were completely alien and their capital more than two months distant. Not understanding Greek political attitudes, they based their control on local tyrants, despite the fact that any form of absolute rule had become an anachronism to the Greeks. It was thus more for political than for economic reasons that the Ionians revolted in 499 BC. When the Naxos expedition failed in 499, its promoter, Aristagoras of Miletus, instigated the Ionian states on the coast of Anatolia to rebel and obtained help from Athens and Eretria. In 498 their forces took and burned Sardis, capital of the satrapy; they were then joined by the Greek states of Cyprus, the Bosporus, and the Hellespont and by the Carians, while Athens withdrew its support. The Persians subdued Cyprus in 496, and the Bosporus and Hellespont by 495. Their naval victory over the Ionians at Lade in the same year, and the capture of Miletus in 493, virtually ended the revolt.

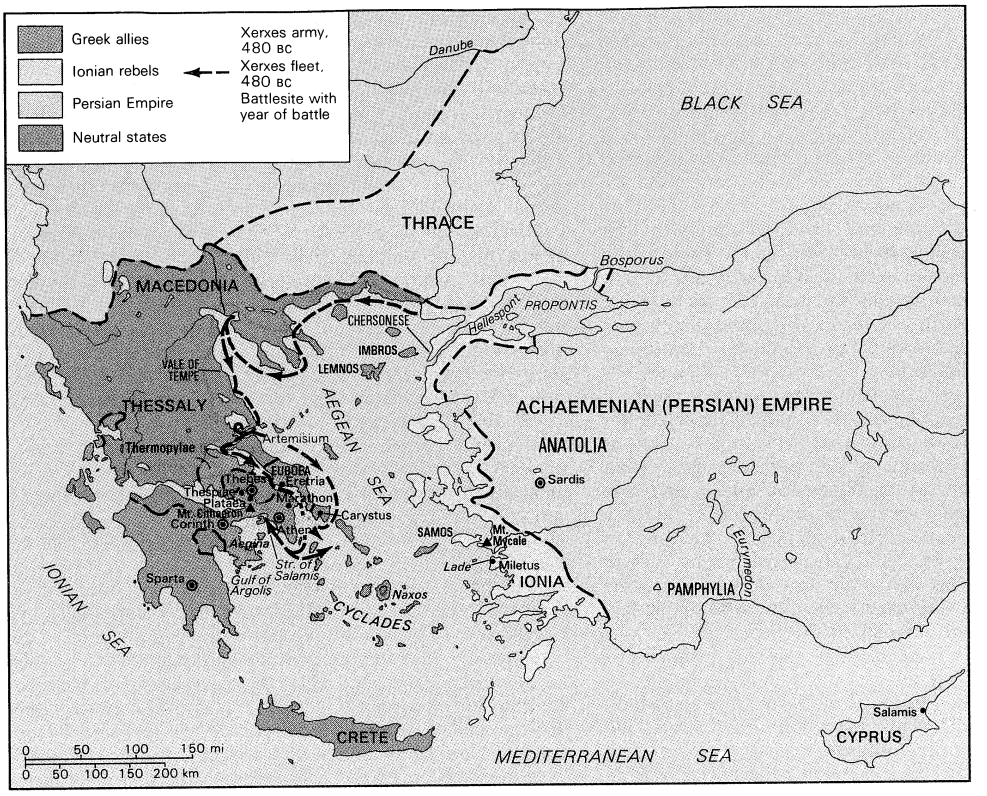
The Ionian revolt was of great value to the Greek cause. It postponed the Persian attack on Greece for a decade and showed the need of close cooperation and strong leadership. The Ionians had indeed created a council of deputies drawn from the individual states and had entrusted to it the direction of strategy, but they had failed to include in the council the Greeks of the Bosporus, Hellespont, and Cyprus, and they had not appointed a commander in chief of the allied forces until the eve of the battle of Lade, when it was already too late.

Persian attack on Eretria and Athens (490 BC). Darius punished the ringleaders in Asia by execution or deportation, but he made a liberal settlement with the states. Democratic governments were permitted, a moderate rate of tribute was imposed, and the states were required to submit their disputes to arbitration. This was politic, since Darius hoped to use the Ionian fleet against Greece. In Europe, his son-in-law Mardonius reestablished Persian rule in Thrace and reduced Macedonia, while his envoys visited the free Greek states to ask for "earth and water," the tokens of submission to Persia. In 491 Eretria and Athens were well informed that an attack by sea was impending.

Before the Ionian revolt Sparta and Athens had been at war, but the Persian threat brought them closer together. In 491 the Spartans tried to prevent their ally Aegina from joining Persia, and in 490 Athens attacked Aegina (the chronology is disputed). Athenian policy toward Persia had vacillated before and during the Ionian revolt, but the will to resist was now strengthened by the Persian support of the exiled tyrant Hippias and by the advice of Miltiades, previously a ruler in the Thracian Chersonese, who had returned home with knowledge of Persian tactics. Even so, Sparta and Athens had no plans for united action when the Persian force, perhaps comprising about 25,000 fighting men, sailed across the Aegean Sea, landed on Euboea, and captured Carystus and Eretria. In September 490 the Persian army landed unopposed on the Plain of Marathon in northeast Attica, whence the lines of supply with Euboea and the east were easy and secure. The speed and the initiative of the Persians found Athens still isolated.

Battle of Marathon

The Athenian army was at Athens, prepared to repel any landing in its vicinity, and the small fleet was ready to attack any Persian convoy heading for Aegina. When news came from Marathon, a runner was dispatched to inform Sparta and the assembly decided on the proposal of Miltiades to send the heavy-armed hoplite army (infantry) to the foothills above Marathon. The decision was wise; for the alternative, to stay and defend Athens, would have cut Athens off from Sparta by land and sea, exposing the city to blockade. At Marathon, however, the ten generals (of whom each held operational command for one day, according to Herodotus) and Callimachus, the "polemarch" or nominal commander in chief, had to choose between attack and delay. Miltiades' advice prevailed: to attack as soon as opportunity offered. First the Athenians advanced their position to within a mile of the enemy by felling trees and making obstacles against the dreaded Persian cavalry. Then the opportunity came. Before dawn some Ionian deserters reported "The cavalry are away." Miltiades, who chanced to hold the operational command, attacked at dawn. With a thin centre and strengthened wings the line of about 10,000 Athenians and 1,000 Plataeans charged the enemy infantry before the cavalry could return. The Greek wings defeated the Persians and wheeled inward to attack the Persian centre, which had driven the Greek centre back. The longer spear and heavier armour of the bronze-clad Greek infantryman prevailed over the Persian with his short spear, wicker shield, and padded cloth-



Greece during the Greco-Persian Wars.
Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York

ing. The rout was complete. According to Herodotus, the Greeks lost 192 men, the Persians 6,400. The majority escaped to the fleet, which sailed at once, hoping to surprise Athens, but the Athenians—by a forced march—arrived that evening to defend the city. The Persians then departed. A Spartan force, which had been delayed by religious observances at Sparta, arrived at Marathon too late to take part in the battle.

In the decade following Marathon there were keen political rivalries at Athens. Miltiades, the hero of Marathon, was discredited in the following year when an expedition against Paros, which he had advocated and led, failed miserably. The death sentence was commuted to a heavy fine, but he died shortly afterward. There followed a quick succession of ostracisms; in 487 and the next two years the victims were friends of the Peisistratids, who were suspected of treason at the time of Marathon. Later, other issues became more important: the most critical was a struggle between Aristides and Themistocles, probably over naval policy, which resulted in the ostracism of Aristides in 482. This struggle was associated closely with a dramatic decision that, more than any other, influenced the course of Greek history in the 5th century. A particularly rich vein of silver had been discovered in the mining area of Laurium, and it was at first intended to distribute this windfall among the citizens. Themistocles, fully aware of the growing danger from Persia, persuaded the people to use the money to build triremes. His new Athenian navy was to prove one of the decisive factors in the repulse of the Persians.

Expedition of Xerxes (480–479 BC). The Persian failure was followed by a full-scale invasion. It was delayed by a revolt in Egypt and the death of Darius until 480, when Xerxes crossed the Hellespont in late spring with a vast army and a large fleet. The advance was slow and the fleet had to provision the army. The Greeks therefore had ample time to make preparations. The problem of uniting the 30 states that had the will to resist Persia was solved by Sparta, which held a congress of delegates and formed a general alliance. The states agreed to stop all wars among themselves and conferred the command by land and sea on Sparta. The congress met regularly, each state having one vote, and decisions by the majority were binding on all members. It possessed recruiting, diplomatic, and judicial powers. In the field the commander in chief, who was nominated by Sparta, consulted the commanders of the national contingents but made his own decisions.

Thus the Greek congress was a highly centralized and efficient organization for allied action. Its chief strength on land lay in the Spartans and their allies; at sea it lay in the Athenians, ably led by Themistocles, who had increased the fleet to 200 ships. A coordinated defense on both elements was now possible. It lay with Sparta to choose the time and place for applying the relatively small but excellent forces of the Greek congress.

The first decision, to hold the narrow Vale of Tempe between Macedonia and Thessaly, was abandoned when it was realized that the position could easily be turned. On news of this the Athenians voted, on the proposal of Themistocles, to entrust themselves to "the wooden wall" of their ships in accordance with an utterance of the Delphic oracle, and plans were made for evacuating the noncombatants. Next the Greeks occupied the still narrower pass of Thermopylae with 6,000 or 7,000 hoplites and stationed 271 ships at Artemisium in northern Euboea. The positions were linked by communication between the Spartan commanders, King Leonidas at Thermopylae and Eurybiades at Artemisium, who intended to thwart the Persian forces. Meanwhile, Xerxes was advancing slowly. He made no use of separate columns, and his fleet suffered heavy losses in a storm when it was convoying supply ships along the coast. It was already August when Xerxes began the operations, which extended over three days. Thermopylae

On the first day he sent a detachment of 200 ships, unseen by the Greeks, to sail round Euboea and close the narrows of the Euboean Channel; and he also attacked with his best infantry at Thermopylae, where the Greeks inflicted heavy casualties. During the afternoon the Greek fleet, having learned about the Persian detachment from a deserter, engaged the main Persian fleet with some success. The Greeks intended to sail south that night and destroy the detachment next day, but a tremendous storm kept the Greeks at Artemisium and wrecked the 200 Persian ships off south Euboea. On the second day news of the Persian disaster was brought up by a reinforcing squadron of 53 Athenian ships. Xerxes attacked again with no success at

Thermopylae, and the Greeks sank some Cilician vessels off Artemisium. That evening a Greek traitor, Ephialtes, offered to guide the Persians along a mountain path and turn the position at Thermopylae. The Persians' best infantry, called the "Immortals," were entrusted to him. At dawn on the third day they began to descend toward the plain behind the Greek position. Leonidas retained the troops of Sparta, Thespiae, and Thebes and sent the remainder south. He then advanced. Except for the Thebans, who surrendered, he and his men fought to the death. Meanwhile the Persian fleet attacked at noon. Both sides suffered heavy losses and the Greeks realized that they could only succeed in narrower waters. That evening, when the fall of Thermopylae was known, the Greek fleet withdrew down the Euboean Channel and took station in the narrow Strait of Salamis.

In September, Xerxes, joined by many Greeks north of Attica, burned Athens. The city was almost deserted, for the evacuation had been completed. The Greek congress decided to fortify the isthmus and keep the fleet forward at Salamis. This decision caused dissension among the ship captains. Many wished to retire to the Argolic Gulf. As a stratagem, Themistocles informed Xerxes of their desire; Xerxes, who saw the end of the campaigning season close at hand, sent 200 ships that night to cut the Greek line of retreat and posted the main fleet, numbering probably 1,207 ships off the eastern exit of the Straits of Salamis. During the night the Greeks learned of his dispositions and intentions. Putting to sea at dawn they feigned a retreat, actually sending a detachment northward to look out for the 200 Persian ships, and their manoeuvres led the enemy to advance incautiously into the narrow waters where superior numbers were of little effect. Within the narrows the Greek ships, stoutly built for ramming, had room to manoeuvre against the congested stream of Persian ships, which, designed for boarding tactics, proved less handy under oar and fell foul of one another. The result was a complete triumph for Greek seamanship. The Persians fled in confusion. Soon afterward their fleet, still superior in numbers but not in morale, set sail for Asia.

That winter, while Xerxes departed to Asia, a large army wintered in Thessaly under the command of Mardonius. By skillful diplomacy he drew the Greeks forward in the summer of 479 to the northern foothills of Mt. Cithaeron near Plataea, where difficulties of supply forced the Greek army of 110,000 men to withdraw during the night. The withdrawal was disorderly and dawn found the army scattered. Mardonius at once attacked a group of 11,500 Spartan and Tegean hoplites who had halted on hilly ground. Their commander, Pausanias, undismayed by the swarms of Persian infantry, led his men downhill in close formation, charged at the double, and overwhelmed the enemy. When the Athenians came up after defeating the Thebans, the Greeks stormed the camp and the survivors of the Persian army fled. Meanwhile, the Greek fleet had passed to the offensive at Mycale on the Asiatic coast opposite Samos. The Persians refused battle, beached their ships, and joined a large supporting army, but the Spartan king Leotychidas landed his men farther north and attacked with complete success. The victories of Plataea and Mycale ended the Persian invasion. (N.G.L.H.)

Plataea and Mycale

The western Greeks. The Greeks of the west had not been able to help in the defense of Greece because they had troubles of their own. While the Phoenicians of the east formed the backbone of Xerxes' fleet, the Phoenicians of Carthage were hoping to drive the Greeks of Sicily out of the island. In the late 8th and early 7th centuries the Greeks had established colonies in the east of the island and the Phoenicians in the west at Motya, Solus, and Panormus. For more than 100 years no hostilities are recorded between the two peoples, but friction developed in the 6th century when Carthage—which by then was the strongest and richest Phoenician city in the west and which had made an alliance with the Etruscans—attempted to exclude Greeks from the western Mediterranean.

By 500 it was merely a question of time before a serious crisis arose. It came in the form of a quarrel between tyrants. Gelon of Syracuse had marriage alliances with Theron of Acragas. Their main enemy was Anaxilas of Rhegium, on the toe of Italy. He had seized Zancle on the opposite Sicilian coast and was allied with Terillus, tyrant of Himera, who was on good terms with the Phoenicians. Theron drove out Terillus and made his son ruler of Himera. Terillus appealed to Carthage, and the Carthaginians sent Hamilcar, one of their chief magistrates, with an army that was probably much larger than the Greeks could muster, though Herodotus' figures (300,000 men) are no doubt exaggerated. The Carthaginians landed near Panormus and proceeded to besiege Theron in Himera. Gelon marched across Sicily with 50,000 foot soldiers and 5,000 cavalry, and in a great pitched battle Hamilcar was killed (according to one account he sacrificed himself) and his army routed. The two tyrants were left in possession of a rich booty and thousands of prisoners. As a result of this decisive victory Sicily was to be safe from Carthaginian invaders for more than 70 years. But in 409 the Carthaginians returned and took their revenge. Himera was destroyed and Acragas sacked. Syracuse was saved by another tyrant, Dionysius I, who for 30 years maintained his power by equivocal tactics and saved the east of Sicily from Carthaginian control.

Battle of Himera (480 BC)

Gelon died two years after his victory, and the rule of Syracuse passed to his brother Hieron, who in 474 won a great naval victory off Cumae against the other main enemy of the Greeks in the west, the Etruscans. This victory was of much more than local significance. The Etruscans, who in the 6th century had extended their rule south of the Tiber and into the rich lands of Campania, had lost control of their land communications when they were defeated in a great battle near Aricia by an army of Cumaeans and Latins, which was soon followed by the expulsion of the last Etruscan king from Rome. The destruction of their fleet off Cumae meant that the Etruscans were cut off from Campania and would no longer be a threat to the Greeks of the Bay of Naples and farther south.

Greek offensives against Persia in Asia Minor. After the Battle of Plataea, the Persian army had abandoned Europe and was not to return. The battle of Mycale was the beginning of another chapter. What was to be the future of the Ionians, many of whom were now in revolt? The Spartans proposed that they should be transferred to the Greek mainland from which they had originally come and be settled on the lands of the states who had medized (collaborated with the Persians). From a military point of view this was good sense. History had shown that without strong help from outside they were subject to the power that controlled Asia Minor. The Ionians, however, had no wish to leave their good lands; they were supported by the Athenians, and the plan was abandoned. It was now near the end of summer, but the Athenians insisted on proceeding to siege Sestos, which was the Persian headquarters in the Chersonese; the Peloponnesians preferred to sail home. By the end of winter Sestos had been starved into surrender and the Chersonese, which was vital to the protection of the grain route from the Euxine, was once again under Athenian control.

Between the battle of Plataea and the summer of 478 the Greeks had decided to continue operations and take the offensive against Persia. The Spartan regent Pausanias was again given the command and led a force of some 100 ships to the island of Cyprus, where Persian garrisons were driven out and the Greeks of Cyprus were encouraged to take control of the island. From Cyprus the fleet sailed up to Byzantium, which commanded one of the two easiest passages from Asia to Europe and the entry to the Euxine. At Byzantium Pausanias became increasingly unpopular with the allies, and there were even rumours that he favoured the Persians, which led to his recall to Sparta. Meanwhile the Athenian commander Aristides had been invited by the Ionians to take over command of the united forces. History had shown that Sparta was ill-adapted to overseas commitments; Athens could claim to be the mother-city of the Ionians and had sent ships to them in the Ionian revolt. Admittedly, they had been withdrawn at the first serious reverse, but it was now a very different Athens. The victory of Marathon, the creation of the largest and most modern navy in Greece, and its

major contribution to the destruction of Persian sea power had increased its confidence and stimulated its ambitions. The Spartan army was still the strongest in Greece, but in 479–478 the Athenians had succeeded in rebuilding their fortifications, despite strong Spartan protests. Considering these developments, then, Aristides probably had a shrewd appreciation of Athens' interests when he persuaded the Athenians to accept the leadership of a new league.

Formation of the Delian League

In the winter of 478–477 Aristides, no doubt in consultation with the leaders of the contingents from Lesbos, Chios, and Samos (who provided the largest of the allied fleets), worked out the basic structure of the new league; all Greeks who were interested were invited to attend an inaugural meeting in the summer of 477 on the island of Delos. Thucydides describes the essentials of the organization:

> The Athenians decided which cities were to contribute money and which ships . . . The *hellēnotamiai* were then first instituted as an Athenian office and they received the *phoros,* which was the name for the money that was collected. The first assessment was four hundred and sixty talents; Delos was their treasury and the meetings were held in the sanctuary. The allies were at first autonomous and policy was discussed at meetings which all shared.

To symbolize the permanence of the association lumps of metal were sunk in the sea and oaths were exchanged. Aristides, on behalf of Athens, and representatives of the allies bound their states by oath "to have the same friends and enemies." From the outset, the influence of Athens, as the freely accepted leader, was paramount. It was understood that it would always provide the commander for military operations and, probably, the chairman of league meetings. It is doubtful whether any formal guarantee of independence was laid down, for the allies were not in a suspicious mood. They realized that without Athenian leadership they would again become a fringe of the Persian Empire, and their feeling was one of undiluted gratitude.

The original alliance included the Cyclades with the exception of Melos and Thera, which had been settled from the Peloponnese; most of the Ionian and Aeolian coastal towns from Miletus to Byzantium; and probably the Dorians to the south, comprising the three cities of Rhodes (Ialysus, Lindus, and Camirus), Cos, and Cnidus. In southwest Asia Minor some Carian communities joined at once, but most of the Carians from the south coast and the interior were not won over until later. The rich cities of the Thracian coast may not have joined the alliance until the middle 470s, but most of the Greek colonies of Chalcidice seem to have followed the example of their mother-cities in Euboea and Andros. By the middle of the century the large islands of Lesbos, Chios, and Samos were the only ship contributors. This was because a much larger number had by then taken the easier option of paying an annual tribute instead.

For some ten years the Athenians continued to lead league forces against the Persians, liberating Greek cities and extending membership of the league. The operations, culminating in a spectacular victory at the mouth of the River Eurymedon (*c.* 466 BC), were carried out under the command of the Athenian general Cimon, and these years could fairly be called the age of Cimon. His series of victories over the national enemy made him the hero of the day, and his family background and his associates gave him a powerful political base. Meanwhile, Themistocles had steadily lost ground. Looking ahead, he advocated anti-Spartan policies, which had little appeal while Cimon, who believed in a dual leadership of Greece, was winning victories in the east.

Cultural developments. The cultural revival of Athens after the Battle of the Eurymedon was at least encouraged, if not led, by aristocratic patronage. The years immediately following the expulsion of the Persians from the Greek mainland had been very bleak. The Acropolis and the lower city had been razed to the ground; public and private buildings had been reduced to rubble and ashes. The main concentration at first must have been on new living quarters and the restoration of the buildings that were essential to the carrying on of public business, but before long architects had new opportunities. When Cimon brought back the bones of the legendary hero Theseus from the island of Scyros a new special shrine was built for them; and from the spoils of the Battle of the Eurymedon he built a new south wall for the Acropolis, enlarging its monumental area.

It is significant that the architects of the Cimonian buildings are now nameless, while the painters whose work they accommodated are remembered. For this was the period of the first blossoming of monumental painting in Greece, a period of bold experiment breaking away from the restricting formulas of the Archaic period. To later generations, Polygnotus was the greatest of these painters. Born in Thasos, the son of a painter, he was probably attracted to Athens by Cimon, who had campaigned in Thracian waters.

Achievements in sculpture

In the sculpture of the period 480–450 BC the same spirit can be seen as in the painting of Polygnotus and his contemporaries. Archaic conventions are discarded and the main concern of the sculptor moves from the surface and linear pattern to a study of solid form. The Archaic smile disappears, and the unreal Archaic symmetry is replaced by more realistic standing figures with the weight of the body unevenly distributed as in nature. A wider and more enterprising range of subjects is chosen, including, toward the end of the period, Myron's bronze figure of an athlete, with body tensed, ready to throw the discus—a bold and difficult subject that was still widely copied in the Roman period. The crown of this generation's achievement was on the Temple of Zeus at Olympia, which was completed by 456. The two pediments, with their contrasting scenes of calm and conflict, and the Labours of Heracles on the metopes, were the finest architectural sculpture that Greece had yet seen.

This was a leaner period for poets at Athens. There was much trivial versifying in honour of Cimon but nothing to match the brilliant circle of poets who had been attracted by the Peisistratids. Simonides of Ceos retained his association with Athens sufficiently long to celebrate its praises in verses on the battles of Artemisium and Salamis, but in 476 he moved to the court of Hieron at Syracuse. Pindar was still in his prime. Before the Persian invasion he had lived and learned at Athens, and he paid a handsome tribute to its share in the preservation of freedom, but the Medism of his native Boeotia strained his feelings toward Athens, and he saw dangers ahead in the development of Athenian policy. His world of aristocratic virtues and privilege was passing away, but there is magic in the rhythm and the imagery of the victory odes for his wealthy patrons in Sicily and Cyrene and for athletes from Aegina and other mainland states.

Athenian tragedy

The most exciting literary development in this period was the growing maturity of tragedy, which was to remain the particular pride of Athens. Aeschylus, who was always regarded later as the greatest of the early tragedians, had fought at Marathon and Salamis and was profoundly moved by what seemed to him the divine salvation of Greece. In his *Persae,* produced in 472, he presented the Persian defeat as a godsent punishment for the hubris of Xerxes, and throughout his career he remained convinced that the divine power who controlled the universe must be fundamentally just. His main concern was to explain the ways of God to man; his characters are cast in the heroic mold, and his language is vivid and colourful. Most of his plays were produced during the Cimonian period, but by the year of his *Oresteia* (first performed in Athens in 458) Cimon had been ostracized, and new democratic forces were transforming Athens.

The Athenian empire. *The changing nature of the Delian League.* Beneath the political calm of the Cimonian period there had been undercurrents that were soon to come to the surface. At the time of the Eurymedon victory there was already a growing change in the attitude of Athens to the alliance. In *c.* 472 Carystus, at the south end of Euboea, which was not an original member of the Delian League, had been forced by military action to join. When, a little later, Naxos wished to secede, it was besieged. Naxos capitulated, and terms were dictated that were incompatible with its independence. It is possible that the revolt of Naxos was instigated by medizing oligarchs, and its reduction may have been in the genuine interest of the league, but when Thasos revolted in 465 the main issue seems to have been Athenian encroachment

on what Thasos regarded as its own economic interests on the mainland. Its navy was defeated and, after a two-year siege, the island surrendered. Thasos had to give up its fleet, pay a war indemnity, and bring annual tribute. Athens took over Thasian mining rights and other interests on the mainland, but the attempt to settle a colony of Athenians and allies at Ennea Hodoi on the Strymon ended disastrously. Penetrating inland, the settlers were defeated by a combined Thracian force, and the colony had to be abandoned.

When Thasos revolted, it had appealed to Sparta and, according to Thucydides, Sparta promised to help by invading Attica. This was one of the first signs that the peaceful coexistence of Athens and Sparta was becoming strained. Hitherto the Spartans had been too preoccupied with their own troubles to worry about the growing strength of Athens. After their resignation of the leadership against Persia, Sparta had sent King Leotychides, the victor of Mycale, to Thessaly to unseat the medizing dynasty of the Aleuads, but the expedition was ineffective and ended with the trial of the King for bribery and his exile from Sparta. There followed a period of political agitation in the Peloponnese. Sparta had first to face an alliance of Argos and Tegea and then an Arcadian alliance that only Mantineia refused to join.

Pausanias, after being acquitted of treachery at Sparta on his first recall, in 478, had sailed out privately and occupied Byzantium, from which he was forcibly expelled by Cimon. He then settled at Colonae in the Troad and from there seems to have made approaches to the Persian satrap at Dascylium. He was recalled to Sparta and, after a surprisingly long delay, incriminating correspondence with Xerxes was alleged to have been found; he was also suspected of stirring up trouble with the helots. Pausanias fled for refuge to the temple of Athena of the Brazen House, from which he was taken when on the point of death from starvation. It is impossible now to recover the true story of Pausanias' last years, but it is probable that the Spartan authorities were more concerned with his possible future plans for Sparta than for his alleged Medism. The Spartans also claimed to have found evidence implicating Themistocles in the designs of Pausanias, and his enemies in Athens brought a charge against him in his absence. He fled from Argos and eventually reached the Persian king, who made him governor of the Greek city of Magnesta on the Maeander. There he died.

When Thasos revolted against Athens in 465 BC, Sparta promised help but was distracted by a serious earthquake, which provided the occasion for the most serious helot revolt since the 7th century. The Spartans were able to subdue the helots in Laconia, but the main strength of the revolt was beyond Mt. Taygetus in Messenia. There the helots took refuge on the very defensible slopes of Mt. Ithome, from which the Spartans could not dislodge them. At this point they called for help to their erstwhile allies in the war against Persia and, among them, to Athens. In the Athenian assembly, Cimon, well known for his sympathy with Sparta, urged support, but there was strong opposition from new men who were working for a more democratic Athens free from any association with Sparta. Cimon narrowly won the vote and led a strong hoplite force to Sparta, but the Spartans found the attitude of the Athenians suspicious and dismissed them, alone of their allies.

Revolt of the helots and the dismissal of the Athenian force by Sparta

Whatever reasons the Spartans may have had for dismissing the Athenians, the results of their action were far-reaching. Cimon was discredited and in 461 he was ostracized. His main opponent, Ephialtes, was able to carry a series of laws that vitally changed the character of the Athenian democracy. The aspect of these reforms that is most emphasized in tradition is the curtailment of the power of the Areopagus, or Areopagite council, which the archons joined at the end of their year of office, but neither Aristotle nor any of his successors was able to define specifically what powers had been lost. In the 6th century the Areopagus had been the main guardian of the constitution, but the reforms of Cleisthenes shifted the balance of power to the Council of Five Hundred and the Assembly, and when, in 487, the lot was introduced in the election of archons, the prestige of the Areopagus probably declined. The Areopagus must previously have been exercising powers that were considered undemocratic. The definition of the very limited cases, all of a religious nature, that they were henceforward entitled to try implies that they had been acting more widely as a judicial court. The emphasis in the new, more developed democracy of Athens on the people's control of the executive suggests that magistrates had previously been influenced by the Areopagus. New powers were now explicitly given to the people. The jury courts were reorganized on a popular basis, and court decisions no longer rested with the magistrates but with the jurors, who were paid for their services, a principle that was later applied to the Council of Five Hundred and to all offices of state; no citizen was to be barred from state service by poverty. All officials had to undergo an examination in a people's court at the end of their year of office, and public auditors were annually appointed to examine accounts. It is significant that from the 450s important decrees of the Assembly and the summary accounts of public works were normally inscribed on stone and set up on the Acropolis or in some other public place where all could see them.

The dismissal of Cimon's force of Athenians by the Spartans led to a reversal of foreign policy. The immediate Athenian reaction was to conclude an alliance with Argos, Sparta's traditional enemy, and with Thessaly, with which Athens had been allied in the 6th century. A dangerous situation became more tense when Megara, a member of the Peloponnesian League, under pressure from its neighbour, Corinth, was accepted into alliance by Athens.

Developments in the west. The 460s were a period of change in Sicily also. For two generations Sicilian history had been dominated by the tyrants of Gela, Acragas, and Syracuse—their cities had enjoyed a period of great prosperity, based primarily on military strength. The tyrant Gelon, who came to power in 491, had made Syracuse the strongest and richest city in Sicily. Marriage alliances linked him with Theron, who had become tyrant of Acragas in *c.* 488, and he proceeded to build up an army and navy to extend his power. He built a wall to defend the extension of Syracuse on the mainland and increased the population at the expense of smaller cities.

Syracuse under the tyrant Gelon

With the exception of a northern pocket around Messana, Gelon came to control the east of Sicily; and by 480 his army was sufficiently powerful and strong to prove decisive at the Battle of Himera when, with his ally Theron of Acragas, he defeated a force of Carthaginian invaders led by Hamilcar. The rich booty from this victory enabled the two cities to build impressive new temples, and the magnificence of the two tyrants' courts became a byword throughout the Greek world; it was reflected in their chariot victories at Olympia and in their dedications at Olympia and Delphi. Gelon died two years after the great victory. Hieron succeeded to the tyranny of Syracuse, being replaced at Gela by the third of the brothers, Polyzelus. It is probable that Hieron was less popular than Gelon and more oppressive in his rule. In his relations with the other Greek cities of east Sicily he was as ruthless as his brother. He drove out the populations of Naxos and Catana and forced them to settle in Leontini, and near Catana he established his own new city of Aetna with settlers from the Peloponnese and Syracuse. His relations with his brother Polyzelus, however, became strained, and when Polyzelus took refuge with Theron at Acragas, the two allies came near to war. When Theron died in 472 his son Thrasydaeus seems to have thought that the best way to counter internal opposition was by war. He assembled a large army, but Hieron struck first, and in a hard-fought battle, in which 4,000 men of Acragas and 2,000 Syracusans are said to have been killed, Thrasydaeus was defeated and was forced to leave Acragas. He took refuge in mainland Megara but was there condemned to death. Hieron made peace and, though Acragas was to make another challenge in 446, Syracuse never lost the primacy in Sicily that Gelon had established.

Hieron was not merely a man of war. Greek literature benefitted greatly from his generosity and taste. Poets need patronage, and Hieron followed the precedent of earlier tyrants, the Cypselids at Corinth, the Peisistratids at Athens, and Polycrates at Samos, on a more magnificent

The tyrant Hieron as patron of the arts

scale. In the desolation following the Persian sack, Athens was in no position to attract a literary circle. Pindar was glad to visit Hieron's court and write odes to celebrate his victories at Olympia. Simonides of Ceos, who had commemorated the victories over Persia, came to Syracuse in 476 and stayed there as a court poet; Bacchylides, his nephew, joined him. Aeschylus was also attracted and gave a special production of his *Persae* in the theatre that Hieron had built; his *Aetnaeae* was no doubt intended to commemorate Hieron's new foundation of Aetna.

Hieron died in 467 at his own Aetna. The youngest of the four brothers, Thrasybulus, who succeeded him, had neither his ability nor the support to maintain his position, and was before long forced to withdraw to Locri Epizephyrii.

The Syracusans celebrated the end of their tyranny by establishing an annual Freedom Festival and setting up a colossal statue of Zeus the Liberator. There seems to have been no serious opposition to democracy, and a democratic crusade was led against the few remaining tyrants in the island. The fighting in Syracuse had repercussions in many other cities, but finally a general agreement was reached that those who had been uprooted by the tyrants should be restored to their own cities and that the mercenaries were to settle in Messana.

Of relations between the Greek cities of south Italy very little is known. After the destruction of Sybaris in 510, Locri Epizephyrii was the strongest of the states, but the archaeological evidence suggests a fairly widespread prosperity. The descendants of the refugees from Sybaris attempted near the mid-century to resettle their old home, but they lacked the strength to hold their own and appealed to the mainland for help. The Athenian settlers who came to join them became a source of friction rather than of strength, and the result was the planting by Athens of a new Panhellenic colony on a new site at Thurii, in 443.

In the first half of the 5th century there was a considerable loss of life in both Sicily and Italy in battles between Greeks and between Greeks and barbarians; but in this same period the achievement of the western Greeks in peaceful fields was very remarkable and comparable with that of the east Greeks in the 6th century. Following the incorporation of the eastern Greeks in the Persian Empire, refugees from Ionia probably played an important part in stimulating the western Greeks. When the Persians came, a large body of Phocaeans sailed to Corsica and, when driven out, settled at Elea (Velia) on the west coast of Italy. Xenophanes of Colophon, a vigorous free-thinker, after much wandering, also settled in Italy. A little later Pythagoras left Samos for Croton, where his compound of philosophy and religion made a deep impression.

During this period, the west was the leader in intellectual speculation. Rival schools flourished. Pythagoras was a considerable mathematician and seems to have thought that numbers were the ultimate reality, but he was also strongly influenced by mystery religions. It is from this source that he based his belief in transmigration of souls, one of the few doctrines that can be certainly attributed to him.

Pythagoras and Parmenides

His followers, observing a rule of life that combined principles drawn from philosophy with a religious mysticism, exercised a strong influence on politics. Their stern morality was effective at first, especially at Croton, where Pythagoreanism had evolved toward the end of the 6th century. The movement spread to Metapontum and other cities, but by the middle of the 5th century the emphasis on moral reform savoured too much of oligarchy, and there was a wholesale reaction against Pythagoreans.

Parmenides of Elea was more important in the history of philosophy. The Milesian speculators had tried to explain the nature of the physical universe, but Parmenides turned from physics to metaphysics. The basic question that puzzled him was what lies behind what one says or thinks: what is the corresponding reality? He dismissed sense perceptions and by involved but logical deductive arguments concluded that reality was a unity—indivisible and unchanging. He established a considerable following, among whom Zeno was his most persuasive supporter, and his arguments, in verse, greatly influenced Plato.

Empedocles of Acragas was a colourful contemporary who, like Parmenides, explained his views in verse. He rejected Parmenides' approach and defended the senses, explaining their distortions in physical terms. He was closer to the Milesians; but instead of searching for a single substance to explain the world, he found four basic elements—earth, air, fire, and water—which, under the changing pressures of love and strife, took varying forms. But there was also a strain of mysticism in his teaching, for, like the Pythagoreans, he believed in the transmigration of soul and would not eat the flesh of any animal.

The poets of the 6th century BC, Stesichorus of Himera and Ibycus of Rhegium, had no distinguished successors, although Epicharmus established comedy in Syracuse independently of Athens. None of his plays survive, but in later tradition he was better known than his Athenian contemporaries. With democracy came an increase in the importance of public speaking; Corax and his pupil Tisias, both Syracusans, were regarded as the pioneers in the study and teaching of the art of rhetoric.

Visual arts of the western Greeks

Unlike the Greeks of the mainland, the western Greeks had no easily accessible marble suitable for building, but their shelly limestone is an attractive material in texture and colour, and their temples in a severe Doric style are impressive for their size and proportions; the temple of Olympian Zeus at Acragas was considerably larger than the Athenian Parthenon. The lack of marble helps to explain why little is heard of west Greek sculpture, and it is significant that Pythagoras (not to be identified with the philosopher), the only sculptor of the west who was later regarded as one of the great masters, came from Samos before settling at Rhegium. But there was one field in which the west Greeks were supreme. The coin engravers of Sicily deserve a place in any anthology of Greek art. There are few coins more beautiful than the Demareteion, a ten-drachma piece, possibly issued to commemorate the Battle of Himera, and named for Gelon's wife.

The first Peloponnesian War (460–445 BC). The ten years that followed the dismissal of the Athenians from Ithome by Sparta (462 BC) and the ostracism of Cimon witnessed a dramatic rise and fall in Athenian fortunes as the Athenians consolidated their democratic reforms and fought on two fronts—against Peloponnesians and Persians. In 460, 200 Athenian and allied ships sailed to Cyprus, which, since the expedition led by Pausanias in 478, had been partly recovered by the Persians. This fleet, however, was soon diverted to what seemed an even more promising opportunity. For Inaros, a Libyan prince who had raised a revolt in the Nile Delta soon after the accession of the new Persian king, appealed to Athens for support. The restoration of Egyptian independence would have been a more serious blow to Persia than the liberation of Cyprus. The rich grain resources of Egypt may also have influenced the Athenian decision to intervene. At first all went well. The Greeks defeated the small Persian fleet and won control of the Nile. The remaining Persian troops fell back on Memphis at the head of the Delta and were besieged; but, though the Greeks quickly occupied most of the town, a determined force of Persians held out stubbornly in the White Fort.

Meanwhile, hostilities had broken out in Greece. The brief surviving record does not explain the circumstances that led to the fighting. The narrative begins with an Athenian attempt to take Halieis, a port on the Peloponnesian coast beyond Troezen, which could have proved a useful naval base for the Athenian fleet for raids on the Peloponnese and for strengthening communication with Argos; but forces from Corinth and Epidaurus drove the Athenians off, though the attempt to destroy the fleet on its return journey resulted in an Athenian victory off the island of Cecryphaleia. At this point Athens' old rival Aegina determined to put out its full strength, but the Athenians, reinforced by league allies, won a decisive victory in a great naval battle. The Aeginetans were driven from the sea and, after a long siege, their city was captured. Aegina was forced to become a member of the Delian League and pay the high annual tribute of 30 talents, the highest of the assessments, shared only by Thasos. During the siege, Corinth had tried to

relieve the pressure on Aegina by invading the Megarid but met with no success.

Direct conflict between Sparta and Athens (457 BC)

Hitherto the fighting had been between Athens and the members of the Peloponnesian League near the Isthmus. In 457 the Spartans intervened. Until 460 (and possibly longer) Sparta had been preoccupied with the resistance of the helots on Mt. Ithome, but they had now surrendered on condition that they could freely leave the Peloponnese. The Athenians gave them a home in Naupactus on the north shore of the Corinthian Gulf, which they had recently captured. In this strategic site, commanding the entrance to the gulf which Corinth regarded as its own, the helots were to repay abundantly the generosity of Athens. The desertion of Megara from the Peloponnesian League and its subsequent alliance with Athens had severely limited the range of Spartan action, for the Athenians had built a wall between Megara and Nisaea, effectively barring the only practical route for a Peloponnesian army invading Attica. The Athenians had also strengthened their position by building long walls from the city to the Piraeus and to Phaleron, within which the population of Attica could take refuge in the event of invasion. To demonstrate their power outside the Peloponnese, the Spartans took a strong force of some 11,500 Peloponnesians across the Corinthian Gulf. Their immediate objective was to defend the small state of Doris, a task that was soon accomplished. They now found that their way home was barred. The Athenians had sent a fleet into the Corinthian Gulf and had closed the mountain route over the Isthmus. The Spartans moved into Boeotia and encamped at Tanagra, threatening Attica and negotiating with extreme oligarchs in Athens, who were anxious to overthrow the democracy. The Athenians marched out, and the pitched battle that followed was evenly fought, but the Thessalian cavalry deserted and the Athenians had to admit defeat. The Spartans, however, did not feel strong enough to attack Athens itself and were content to return through the mountain passes of the Isthmus. Athenian morale was not badly shaken and two months later they returned under their general Myronides and won a decisive victory over the Boeotians at Oenophyta (457 BC). This victory gave them control over all Boeotia, with the possible exception of Thebes, and their control was soon extended over Phocis and Locris. They could now turn to the offensive against the Spartans; the Spartan dockyards at Gythium were burnt and Chalcis on the Gulf of Calydon was captured. This marked the high tide of Athenian success.

The tide first turned in the east. The Persian king had tried to relieve the pressure on Egypt by bribing the Spartans to act more vigorously against Athens. When this failed, a large army and navy were sent to crush the revolt. A squadron of 50 ships, sent out to relieve the Greek fleet or part of it, entered the Nile and was largely destroyed. The Greeks were driven from Memphis and fell back on Prosopitis, an island between two branches of the Nile. After 18 months the Persians dried up the canal linking the two branches, and the Greek position became untenable. Under an armistice the survivors were allowed to cross the western desert to Cyrene. The loss in men and ships was serious, and the effect on Athenian expansion decisive. During the early 450s Athens had aimed at bringing into the league Cyprus and all cities in the eastern Mediterranean that had Greek roots. The collapse in Egypt meant the abandonment of Cyprus and the eastern Mediterranean. In Greece itself the offensive against the Peloponnese was halted. Over the next three years the Athenians concentrated on building up their strength, but before a new expedition sailed east they wisely decided to secure themselves against Sparta. No one could have been a more acceptable mediator than Cimon, who had always advocated friendship with Sparta. Cimon was, therefore, specially recalled, shortly before his ten years' ostracism was due to expire, and he succeeded in negotiating a five years' truce early in 451.

Cimon could now resume command of league forces against Persia, and in 451 (or possibly 450) he sailed east with 200 ships. His main objective was the recovery of Cyprus, but he detached 60 ships to Egypt, where revolt was still smoldering. In Cyprus the campaign started well, but Cimon died while besieging Citium, the strongest Phoenician city in the island, and his colleagues in the command decided to take the fleet home. The Greek withdrawal from Cyprus marked the end of fighting between the Delian League and Persia. Both sides could now leave the struggle with some dignity, and both had good reason to do so. The Athenians realized that they had overstrained their resources and felt nervous for the stability of the league; the Persians had recovered control over the eastern Mediterranean and Egypt, although they were faced with a revolt in Syria. It is probable that a formal peace, the so-called Peace of Callais, was now made, in which the Athenians abandoned their ambitions in the eastern Mediterranean and the Persians agreed not to attack the members of the Delian League and to keep their fleet out of the Aegean; but there is no firm evidence, and the peace remains controversial. It is certain, however, that Athens decided to terminate hostilities.

Cimon's expedition to Cyprus

The years that followed the end of fighting against Persia were the most critical period for Athens, both at home and in relations with the allies, since the invasion of Xerxes. The most important problem was the future of the Delian League now that Persia could no longer be regarded as an active enemy. Athens was determined to maintain the league and did not hesitate to make this clear to the members. The Athenians' relations with the allies had, by 465, changed considerably, and they had used force to put down revolts. In the 450s, some of the cities were called upon to send ships or troops to fight against the Peloponnesians—something that had certainly not been anticipated earlier. Athens removed the league treasury from Delos to Athens. Henceforth the annual tribute was to be brought to Athens in early spring at the time of the Dionysia. The transfer of the treasury could be justified when it was made, for the Persian victory in Egypt might have encouraged the Phoenician fleet to sail into the Aegean and raid Delos. But however honest the motive may have been, the transfer made it much easier for the Athenians to use the money for their own purposes.

After the death of Cimon, Thucydides, son of Melesias (not the historian), and probably Cimon's brother-in-law, set out to organize into a coherent party the extreme oligarchs and all those who might have an interest in reversing the tendencies of the democratic leadership. When the Assembly decided to use the tribute from the allies to build Athena's temple, Thucydides tried to persuade the people to repudiate the immoral use of the allies' money:

> Greece appears to be the victim of monstrous violence and manifest tyranny when she sees that with the money contributed under compulsion for the war we are gilding our city like a wanton woman, adorning her with extravagant stones, statues and thousand-talent temples.

But the pride of a new temple and the prospect of employment that it offered were sufficient to dull any tender consciences. Thucydides continued to attack his democratic opponents, but Athenian policies were not significantly modified.

Difficulties within the Delian League

Athens also had trouble from the allies. The tribute lists of the years from 450 to 446 show that this was an unstable period. Many cities paid late in the year; others made incomplete payments and some, including probably Aegina, did not pay at all. It was probably in this period that Colophon revolted and was suppressed. There had been trouble in Erythrae and Miletus before 450, and the Milesian oligarchs, with whom the Athenians at first tried to cooperate, now carried out a purge of their opposition. But by 440 democrats were in control of Miletus; they had almost certainly been installed in power by the Athenians.

Two important decrees that illustrate the new imperial tone probably date from these years. The first required all the allies to use Athenian coins, weights, and measures and banned all minting of silver: now that the tribute was to be spent mainly in Athens, crews, contractors, and workers naturally wanted to be paid in Athenian coin. But the move was also an open expression of empire. Another attempt to bind the cities together in an empire was alluded to in a second decree, the main purpose of which was to tighten up the collection of tribute but which referred in passing to a more important decree, recently passed, requiring all the allies to bring to the festival of the Great Panathenaea—held every four years—a cow and a suit of armour: colonies were traditionally expected to

recognize in such ways the main festivals of their mother-cities. Athens also hoped to encourage loyalty to the empire by spreading the cult of "Athena, queen of Athens" in the cities.

Another important feature of this period was the settlement overseas of Athenian citizens of the two poorest classes. This policy of establishing what were called "cleruchies" in the territories of allied cities was one of the main grievances still remembered against Athens in the 4th century, although not all such settlements were unpopular. Some cleruchies were intended as garrisons and settled as penalties for anti-Athenian attitudes or actions.

Periclean democracy and imperialism. By the various measures taken between 450 and 446, Athens had succeeded in converting a league into an empire. The opposition in the Assembly led by Thucydides had been active but ineffective, and allied discontent had been met by firmness. There was no attempt to disguise the change of status, and the Athenians freely admitted that they no longer led autonomous allies. The strength of the new Athenian position was fully confirmed when Athens successfully weathered a succession of serious crises in *c.* 446, the year when the five years' truce with Sparta was due to expire. In spring 446 or summer 447 the Athenian general Tolmides underestimated the seriousness of a rising in Boeotia. He was surprised and defeated near Coronea, and to save the army the Athenians had to undertake to withdraw entirely from Boeotia, undoing all that had been won at Oenophyta. The Boeotians had been helped by Euboean exiles and revolt now spread to the main cities of Euboea. Pericles, now the leading Athenian politician and general, had taken an army over to the island when he learned that the Megarians had massacred most of their Athenian garrison and that a Peloponnesian army was marching toward Attica. He brought his army back and negotiated a Spartan withdrawal (allegedly by bribery). He could then return to Euboea, where the rebel cities were quickly reduced.

The Thirty Years' Peace

Meanwhile, an Athenian embassy to Sparta had negotiated the terms of a peace which was to last 30 years. The Athenians gave up Nisaea, Troezen, and Achaea. Aegina had also hoped to regain its independence but had to be satisfied with a clause guaranteeing freedom from Athenian interference provided that an annual tribute was paid. Athens' most important gain was the recognition of its empire by Sparta. Both sides undertook not to interfere with the alliance of the other, though neutrals could join either.

During the peace negotiations the Athenians had to draw up new tribute assessments. Some 30 cities (and probably more, for the evidence is not complete) had their assessments reduced, while only three increases are known; the natural inference is that Athens was reacting to the crises from which it had emerged by a policy of concessions. There is no firm evidence of the attitude taken by Thucydides and his followers to the events of these critical years, but in 443 there was sufficient feeling against him to secure his ostracism. With his fall, the party he had built up collapsed. The extreme oligarchs remained implacably hostile to the democracy but were not strong enough to come out into the open; most of the moderates were content to accept the leadership of Pericles, who, in spite of his political views, was an aristocrat.

Pericles, early in his political career, had been personally responsible for the introduction of state pay for state service, first to the jurors and later to the Boule (city council) and all branches of the executive. By the early 440s he was the most influential politician in Athens. He was the main architect of the conversion of league into empire and was almost certainly mainly responsible for ending the fighting against Persia.

Periclean building in Athens

With his firm belief in the need for maintaining a tribute-paying empire, Pericles combined imaginative ideas on the quality of Athenian democracy and culture. He began the transformation of Athenian public architecture, and it was he who formally proposed that the tribute reserve should be used for the rebuilding of Athenian temples. The public buildings of the Cimonian period had been in limestone; the Periclean temples were in marble from the quarries of Mt. Pentelicus. The most impressive of the new buildings were the Parthenon, with its splendid gold and ivory cult statue of Athena, and the new monumental western entrance to the Acropolis. Other fine temples were built in the same generation, among them, one to Hephaestus overlooking the Agora, to Poseidon on the headland of Sunium, and to Nemesis at Rhamnous. The program was not confined to religious buildings. A large concert hall, with its tentlike roof supported by a forest of columns, was built near the theatre on the southeast slope of the Acropolis, and to make the amenities of life available to poor at well as rich, public money was used to build public gymnasia and baths.

Pericles was said to have been a pupil of Anaxagoras, a philosopher who spent the greater part of his active life in Athens. To him he owed his rigorous grounding in the study of the physical universe and a lasting interest in philosophy; together they freed him from the restricting conventions of contemporary religion and superstition.

Pericles was reserved and did not mix freely with the people; but he was a commanding orator, and the Assembly learned to accept his judgment and follow his lead. After the ostracism of Thucydides he was elected general for 15 years in succession. But the real source of his power lay more in his personal authority. The historian Thucydides, in recording the funeral speech he delivered in 431 to commemorate those who died in the first year of the Peloponnesian War, outlines the main qualities of the democracy that Pericles did so much to create. Unlike Sparta, inward-looking and strictly disciplined, Athens was a city wide open to the world. Pericles had himself stiffened the qualifications for citizenship by requiring Athenian parentage on both sides, but foreigners were encouraged to settle in the Piraeus or Athens itself and were better treated and had more profitable opportunities there than in any other Greek city. Athens welcomed them, unskilled as well as skilled, and they played a dominant part in Athenian trade. In the Periclean democracy, public service was not confined to a minority of activists but was the natural duty of all citizens, poor as well as rich, and most offices were filled by the lot. Five hundred men had to be found each year for the Boule, and, as no one could serve more than twice nor be under the age of 30, a large proportion of the citizen body must have served at least once. The Boule reviewed all state business and controlled the executive; and no business could be discussed in the Assembly unless it had been put on the agenda by the Boule. A year in the Boule was an intensive political experience, and for a tenth of the year every member had to be continuously available. The sovereign body was the Assembly, in which every citizen had an equal vote. Policies were normally discussed on the basis of resolutions drafted in the Boule (*probouleumata*), but it was also open to the Boule to make no formal recommendation and merely to bring a matter forward for debate. In either case there was a free vote; anyone could address the meeting; and amendments could be proposed from the floor. Perhaps the most surprising feature of the proceedings of the Assembly was the appointment by lot of the chairman, who changed daily and was chosen from the members of the Boule. That important policies could be discussed at large open-air meetings—at some of which a quorum of 6,000 was required—and result in rational decrees is a great tribute to the tolerance of the average citizen. Thucydides, looking back on what he regarded as the golden age of Pericles said that "it was in name a democracy, but in reality the rule of the first man." This should not obscure the fact that Pericles had to persuade the Assembly and that the Assembly was not an ignorant mob; he could not act arbitrarily without authority. For his actions in office he had to submit to the same public examination as others, and at any time his proposals could be rejected by vote.

Nature of Periclean democracy

Importance of the empire to Athenian prosperity

The high standard of living that the Athenian democracy provided for its citizens depended in no small part on the stability of the empire, the tribute of which paid directly or indirectly for much of Athenian public spending. Without the tribute reserve, Pericles' ambitious building program would have been out of the question, and, without the

annual tribute, Athens would have been unable to afford both state pay for state service and a large and active fleet. There is, however, no evidence that the allies were financially oppressed. There was no general increase in assessments before the Peloponnesian War, and more than half the cities kept the same assessment from 453 to 431 BC. For most cities the tribute was probably a small price to pay for the policing of the seas by the Athenian fleet. Too little is known of the pattern and volume of trading during this period to draw firm conclusions, but it is clear that the Piraeus became the busiest trading port in the Aegean and probably in the Mediterranean. A wide range of goods came in from the Euxine, Phoenicia, Egypt, and Carthage, as well as from the Greek world. There seems, indeed, to have been a considerable growth of trade following the end of hostilities against Persia, and, though Athens was the main beneficiary, the security of shipping benefitted the allies as well. The grievances of the allies were political rather than economic. They had lost their independence and were no longer allies but subjects.

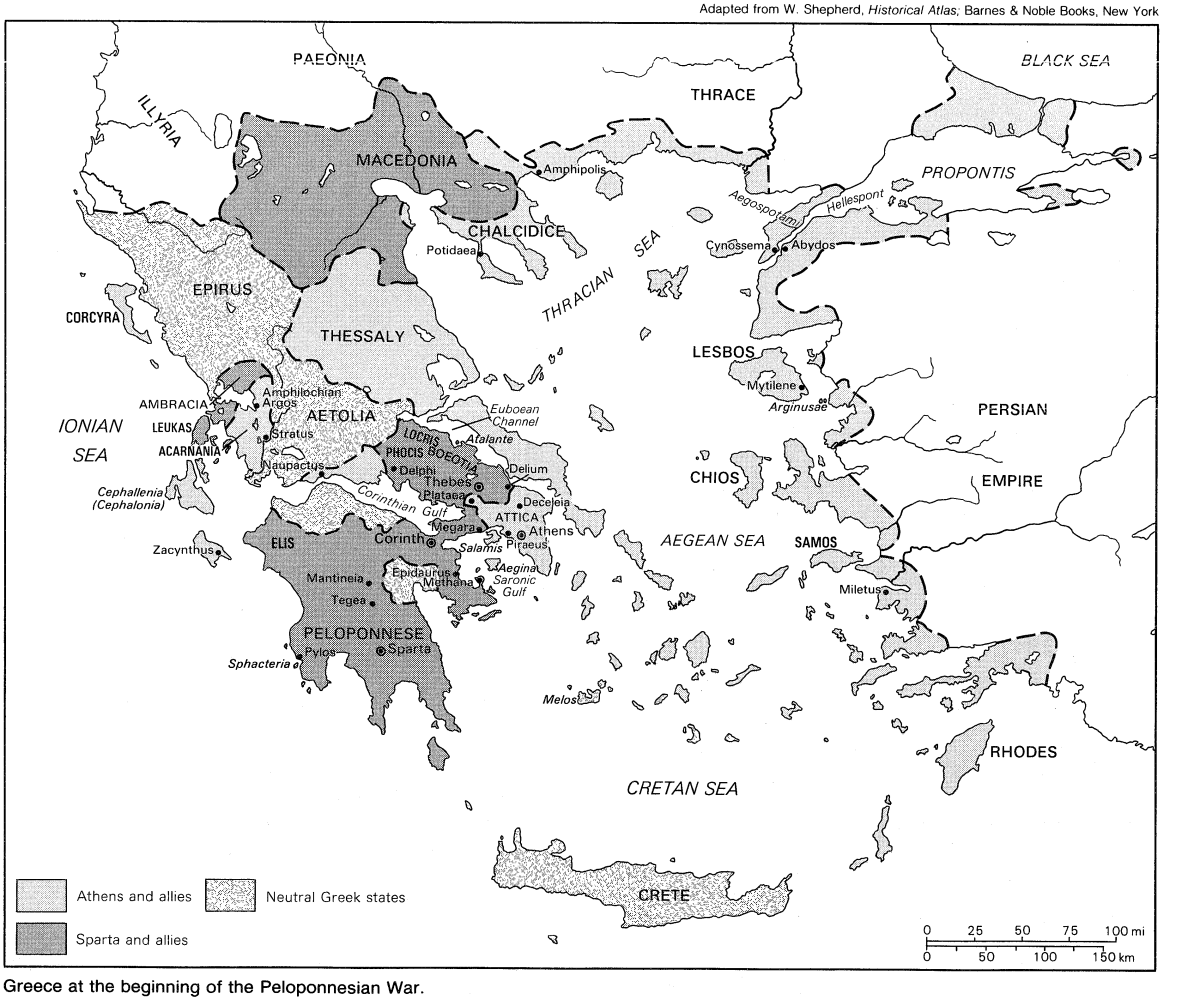
In order to control subject cities, Athenian officials, *archontes,* were installed, sometimes singly, sometimes as many as five; with them went a small security guard, stationed on the local acropolis, capable of dealing with small-scale trouble but not strong enough to crush a serious revolt. Athens also sent out commissioners, *episkopoi,* to cities where trouble had arisen to examine the situation, make recommendations on the spot, and report back. Such officials were widespread by the early 440s. More important was the control exercised by the people's law courts at Athens. All cases involving the major penalties of death, exile, loss of citizen rights, and total confiscation of property had to be referred to Athens. But the most reliable means of control was the encouragement or imposition of democratic governments. It was the oligarchs who most resented Athenian rule; the burden of tribute fell mainly on the wealthy, and the powers of local dynasts were restricted. Most of the revolts for which there is evidence were engineered by oligarchs, and in the settlements that followed power was normally given to the demos. It was inevitable that oligarchs should look elsewhere for support, to Persia or Sparta, and when it came to a choice between rule by their own extreme oligarchs or Athens, the common people would tend to prefer Athenian control.

After the Thirty Years' Peace had been negotiated, the Athenian Empire seemed secure. It was a considerable shock when, in 440, Samos, one of the three islands off the Asiatic shore that still contributed ships, revolted. Fighting had broken out between Samos, which at the time was controlled by oligarchs, and Miletus, where Athens had recently installed a democracy. Priene, which at various times had needed to fight against both for its independence, was the bone of contention. When an Athenian attempt at a political settlement failed, war followed, and the Athenians had to deploy their full naval resources and summon contingents from Chios and Lesbos. After a nine-month siege, Samos had to surrender and was made to give up its fleet, raze its walls, and agree to pay a substantial indemnity in installments. Samos had appealed for help to Sparta, and the Spartans went so far as to summon a meeting of the Peloponnesian League; but the peace had stipulated that neither side should interfere with the other's alliance, and the league voted against war. The Samian oligarchs had also appealed to the Persian satrap at Sardis and had received some mercenaries from him. Of the Athenian allies, Byzantium alone followed Samos but was recovered without serious fighting.

The revolt of Samos

The Great Peloponnesian War (431–404 BC). Five years after the surrender of Samos a train of events started in the northwest that led to the Great Peloponnesian War. Though the record is thin, it seems that in these years Athens was in an expansive mood, a mood likely

Adapted from W. Shepherd, *Historical Atlas;* Barnes & Noble Books, New York



Greece at the beginning of the Peloponnesian War.

to cause misgivings in Sparta. In 436 Athens succeeded in planting a large colony at Ennea Hodoi, now to be renamed Amphipolis, on the Strymon, a splendid base for the exploitation of the rich mineral and timber resources of the region. But Athens could not, without weakening its manpower dangerously, provide more than a nucleus of the colonists; the mixed character of the population, drawn largely from the Greek colonies of the area, was to prove a serious weakness. A little later Pericles himself led a spectacular expedition into the Euxine to demonstrate Athenian strength and their interest in the Greek cities of the south as well as the north coasts. At Sinope the local tyrant was expelled, and 600 Athenians were sent out to strengthen the Greek population.

Athenian intervention between Corinth and Corcyra

By 435 Athenian restlessness was causing general concern, and when fighting broke out between Corinth and its old colony Corcyra there was soon a danger that it would spread much further. For Corinth, humiliated by defeat in a naval battle, began to make intensive preparations for a larger fleet to teach Corcyra a lesson, and Corcyra appealed to Athens. Since Corcyra belonged to neither alliance, the Athenians would not be breaking the terms of the peace if they accepted Corcyra as an ally, although such an action would clearly invite the bitter resentment of Corinth. After an even debate in the Assembly, the Athenians decided to make a defensive alliance: Athens would help Corcyra if it was attacked and was in danger of defeat. Ten ships, shortly followed by a second squadron of 20, were sent out. In the Battle of Sybota that followed, the Athenian ships intervened when the Corcyraeans were in grave danger; the Corinthians withdrew. More bitterness was caused when the Athenians, suspecting trouble from Potidaea in Chalcidice, a Corinthian colony but a member of the Athenian Empire, required the city to pull down its main defensive wall and give hostages. The Potidaeans, after failing to negotiate concessions from Athens, and relying on a Spartan undertaking to invade Attica if Potidaea was attacked, came out in open revolt. It was probably soon after this development that the Athenians charged Megara with cultivating borderland that belonged to the Eleusinian goddesses and with giving a refuge to runaway Athenian slaves. When no satisfaction was given, Pericles carried a decree banning the Megarians from the Athenian Agora (civic centre and market) and from all the ports of the Athenian Empire.

In 432, while an Athenian force was blockading Potidaea, the Corinthians and other members of the Peloponnesian League sent representatives to Sparta to press for action. Sparta was divided; King Archidamus pleaded for caution and delay, but a clear majority preferred the rousing demand for war from the ephor Sthenelaidas. To involve its allies Sparta needed a majority vote at a formal meeting of the Peloponnesian League. In the late summer of 432 the league was committed to war.

Thucydides' view of the war's origin

In his detailed account of these events Thucydides emphasizes the initiative of Corinth in bringing matters to a head, but he saw clearly that the cause of the war went far deeper than the trading interests of Corinth. "The true cause of the war . . . was the growth of Athenian power which compelled Sparta to fight." Other writers laid the main stress on the Megarian decree and on the refusal of Pericles to compromise. To Thucydides the real issue was not the trade rivalry between Athens and Corinth but whether Sparta or Athens should be the leading power in Greece. In the war, which opened in 431, the Spartans were technically the aggressors. They sent ultimatums to Athens over the winter, the terms of which were such that they cannot have expected Athens to accept them. Peace, however, could have been maintained if the Athenians had been more conciliatory. But Pericles had long since decided that Athens could not reach its full development while Sparta controlled the Peloponnese. His vision of Athens included the eclipse of Sparta.

Pericles led Athens into the war deliberately and with full confidence. In spite of the ambitious building program there was still a reserve of some 6,000 talents on the Acropolis which, when combined with annual income, seemed ample to sustain even a long war. The Athenian fleet, with 300 ships, was immensely superior to that of the Peloponnesians in numbers, experience, and skill. The Athenian fleet, Pericles thought, could maintain a firm hold on the empire and attack the enemy's coastline. The Athenian army was weaker in numbers and military reputation than that of the Peloponnesians, but it was sufficient to man the defenses and raid the Megarid. Pericles argued that the army should not be used to fight the Spartans and should avoid pitched battle; Attica must be abandoned and the country population take refuge in Athens, where they would be protected by the city walls and the long walls running to Piraeus. There should be no attempt to expand the empire during the war. It is difficult to see how Sparta could have been crushed by Pericles' strategy, but under continuous pressure the Peloponnesian League might have disintegrated, forcing Sparta to accept an unequal peace, virtually admitting Athenian primacy.

The great weaknesses on the Spartan side were their fleet and their finances. The Spartans still dealt in iron bars for currency, and most of their allies were agricultural states with lean resources. The coastal towns, particularly Corinth, were economically stronger, but they could not afford to raise and pay large fleets. The Peloponnesian army, however, could easily muster 30,000 (against an Athenian limit of 13,000).

The pattern of warfare

The first year of the war established the pattern that was to be expected. The Peloponnesian army, led by King Archidamus, invaded and ravaged Attica, while the Athenians watched helplessly from their walls. Meanwhile, a fleet of 100 Athenian ships sailed around the Peloponnese, raiding selected points. In 430 the strategies were repeated, but the most important event of the year was the arrival in Athens of a plague from the east that swept through the overcrowded city, killing roughly a quarter of the fighting men and an even larger proportion of women, children, and slaves. This had a depressing effect on Athenian morale; an embassy was sent to Sparta to negotiate a settlement and Pericles was dismissed from his command. Sparta, however, would not offer acceptable terms; the Athenian mood recovered, and Pericles was shortly re-elected. But in 429 he died.

The demagogic element in Athens. Pericles' death was followed by a change in the political climate. Among those who had been generals with Pericles there was no outstanding politician, and, as a consequence, policies became less consistent. The most disturbing factor was the emergence of new political leaders who no longer depended for their influence on their prestige as generals or on the distinction of their families but on their shrewd understanding of the common people and their debating skill in the Assembly. Cleon was the first of these new leaders. He first appears in the record in 431, attacking Pericles when the latter would not allow the Athenians to go out and fight. Like his successors, he rose to power as a critic, prosecuting oligarchs in the courts, seizing any opportunities that were offered when magistrates were publicly examined at the end of their year of office, and claiming to be the champion of the common man. Aristophanes caricatures him as a leather seller, as if he were some small-scale retail trader, but Cleon's father had supplied a chorus for a public festival and must therefore have been a man of more than average means. By his background he was better qualified than most of his political rivals to handle financial problems.

The rise of Cleon

Neither Thucydides nor Aristophanes was sympathetic to the new style, but even Thucydides admits that by 427 Cleon was "by far the most persuasive speaker in the Assembly." Thucydides also describes him as "most violent," and from Aristophanes it is known that he had a loud voice "like a mountain torrent." His qualities were demonstrated when Mytilene, which had been in revolt, surrendered in 427 and the Assembly had to decide what penalties to impose. Cleon convinced the people that the right policy was to put all adult males to death and enslave the women and children. At a second meeting, called to reconsider the decision, he stood firm by his original motion and was only narrowly outvoted.

Cleon's influence became even stronger in 425. An Athenian fleet sailing to Sicily had left a force on the promontory of Pylos in Messenia, hoping to encourage Spartan

helots to desert. The Spartans, in alarm, brought back an invading army from Attica; in their operations against Pylos they occupied the neighbouring off-shore island of Sphacteria in order to deny it to the Athenians. But Pylos held out, and the Athenian fleet blockaded Sphacteria. When the blockade proved ineffective Cleon violently attacked the generals in the Assembly, demanding more vigorous action, with the surprising result that he was himself given the command. He took light-armed troops and used the advice of Demosthenes, the most enterprising of the generals. They landed men on the island and finally forced the Spartans to surrender. Now Athens had hostages against further invasions of Attica, and Cleon's followers could introduce an extraordinary assessment in which the tributes of the allies were sharply increased to bring in a total of nearly 1,500 talents, more than three times the prewar figure.

The conduct of the war now became more vigorous. A raid was made on Corinthian territory, and in the summer of 424 the island of Cythera off the south coast of Laconia was captured, providing a useful base for raids on the mainland and for intercepting merchantmen coming to the Peloponnese from Libya and Egypt. In 424, however, the Athenians over-reached themselves. Cleon had been elected general, but it was Demosthenes and Hippocrates who planned the campaigns. Their first objective was to gain control of Megara by a combination of military action and political propaganda, but in this they failed. Demosthenes and Hippocrates then proceeded with an elaborate attempt to regain control of Boeotia. There were to be democratic risings in the cities; Demosthenes was to land with a force at Siphae on the Corinthian Gulf, while Hippocrates led out the main Athenian army and occupied Delium. The plan was too complicated and could not be kept secret. Siphae was garrisoned when Demosthenes arrived; the revolts in the cities misfired; the army from Athens was isolated and, in a pitched battle near Delium, was decisively defeated. At roughly the same time, the Spartan Brasidas and a small force had reached Chalcidice and was attempting to undermine Athenian control of the area. Brasidas met with rather less success than might have been expected, but he succeeded in his most important objective by capturing Amphipolis, a very serious loss to Athens.

The Peace of Nicias. Athenian prospects had now completely lost the promise that the victory at Pylos had brought, but the Spartans were more anxious to recover their prisoners than to continue the fighting in the north. A year's armistice was therefore arranged from April 423; it was intended to pave the way to a permanent peace. But within a few days of the armistice Mende and Scione revolted, and Brasidas supported them. Mende was quickly recovered, but Scione held out into 421. In 422 Cleon was given command in the hope that he would repeat his success of 425; he did indeed start well, winning back Torone and some smaller places, but at Amphipolis he failed. He led out his force to high ground beyond the town in order to reconnoitre, and was caught by a surprise attack while returning. His army was beaten and he himself killed, but Brasidas also fell in the battle. Their deaths made it easier for the Spartans and Athenians to negotiate seriously for peace.

Terms of the peace

The main terms agreed between the two sides were that each should return its prisoners and that Sparta should recover Pylos, Athens Amphipolis. No provision was made for the restoration of Sollium and Anactorium, which the Athenians had taken from Corinth, nor of Nisaea, which they had occupied in 424 when they failed to win Megara. Boeotia was required to hand back Panactum, an Athenian border fort captured after the Battle of Delium. The Spartans had ignored the interests of their strongest allies and, when the terms were submitted to the Peloponnesian League for formal ratification, Corinth, Megara, Elis, and Boeotia all opposed acceptance; but they were outvoted by the smaller states, which had nothing to gain from war; nor were their interests seriously affected by Athenian imperialism.

The peace was named after Nicias, the most influential of the Periclean politicians since the death of Pericles, and the main enemy of Cleon. Nicias lacked Pericles' authority, however, and, though popular, was indecisive and ill-suited to lead Athens firmly to friendly coexistence with Sparta. His peace was doomed from the outset because the war had settled nothing, and too many cities were dissatisfied with the terms—especially Corinth, which had lost most and gained nothing. Internally, neither Sparta nor Athens was stable. When new ephors were appointed in the autumn at Sparta the board of five included two men who were anxious to renew the war. In Athens, Nicias had two powerful opponents who were not prepared to accept his policy. One, Hyperbolus, had inherited the mantle of Cleon; he had the same background, relied on the same following in the common people, and advocated similar policies. Cleon's fortune came from leather, Hyperbolus' from lamps; in Aristophanes' plays they are both warmongers. More dangerous was the young Alcibiades, perhaps the most brilliant Athenian of his generation. His father had died at the Battle of Coronea in 446 when he was only five years old, and he had been brought up in the house of his guardian Pericles. Strikingly handsome, with a quick intelligence and a cheerful contempt for convention, he fascinated women and could lead men. He combined great abilities with a fundamental lack of responsibility and, as Thucydides saw, his private life ruined his public career. When the peace was made he looked for opportunities to discredit Nicias and win the kind of position that Pericles had held. In so doing he roused the bitter hostility of Hyperbolus and his associates because his colourful conduct appealed to the common people. So developed a feud that was to be one of the main causes of Athens' ruin.

Breakdown of the peace: Athens' Peloponnesian alliances

The failure of the Peace of Nicias was soon manifest. Prisoners were exchanged, but Amphipolis was not handed back to Athens because the inhabitants were determined to remain independent, and so Athens continued to hold Pylos. In the uneasy years that followed there was a confusing pattern of negotiations and alliances. First Corinth attempted to build up a third power, winning over Argos, whose Thirty Years' Peace with Sparta was due to expire toward the end of 421, and some of the members of the Peloponnesian League; but an alliance between an oligarchic Corinth and a democratic Argos was bound to be insecure, and Corinth lacked the power and confidence to hold the alliance together. At the time, however, it seemed sufficiently dangerous to persuade Athens and Sparta to join forces in a new alliance to protect themselves. Before the end of 420, however, the influence of Nicias had been seriously eroded and Alcibiades was able to lead the Assembly away from Sparta into an alliance with Argos and two members of the Peloponnesian League that resented Sparta's high-handedness—Mantineia and Elis. This alliance was put to the test in 418 when the two armies faced one another outside Mantineia and Sparta was cut off from Corinth and Boeotia. The battle was hard fought, and at one point a picked corps of Argives broke through the Spartan line; but Spartan discipline finally prevailed, though the defeat did not become a rout.

The combination of Athens and Argos with discontented Peloponnesians near Sparta offered the only practical prospect of defeating Sparta decisively. But the policy failed because Athens' support was halfhearted. Its contribution to the allies' army—only 1,000 hoplites and 300 cavalry—was inadequate, and it was led by Laches and Nicostratus, who had been consistently associated with Nicias; Alcibiades, who originated the policy, had not even been elected general for the year. Elis also was to blame; when it disagreed with the priorities for the campaign, it withdrew 3,000 men.

Having failed in its Peloponnesian policy, Athens turned elsewhere. In 416 it sent a strong force to bring the island of Melos into the empire. Melos had been neutral at the beginning of the war and had resisted when Nicias landed in 426 with a strong force and ravaged the land. Though the Athenians now promised that there would be no victimization if Melos would join the other islands as a tribute-paying member of the Athenian Empire, the Melians stubbornly refused. After a stiff resistance they were forced to surrender; the men of military age were put to death, and the women and children enslaved. This was

the fate that Cleon had advocated for Mytilene and had succeeded in imposing on Scione. But these two cities had revolted; Melos had merely tried to remain independent. From a playful allusion to the "Melian famine" in Aristophanes' *Birds,* produced in 414, it seems that the savage punishment of Melos did not weigh heavily on the Athenian conscience at the time. But Thucydides emphasized its significance, and it was widely remembered against Athens in the 4th century.

The expedition to Sicily and its aftermath. In 415 the Athenians embarked on an ambitious attempt to gain control of Sicily. The occasion was an appeal for help from Segesta in the west of the island, which was under pressure from neighbouring Selinus. Alcibiades saw in a Sicilian expedition the opportunity he needed for spectacular successes, and the people were attracted by the wealth of Sicily. Nicias pleaded for caution but was outvoted. In spite of his strong opposition he was elected to a joint command with Lamachus and Alcibiades. It was a bad compromise. The expedition was the most impressive that had left the Piraeus since the middle of the century. There were nearly 100 battle triremes, including 34 from the allies, with 40 more used as transports, and 5,100 hoplites, of whom only 1,700 were Athenians.

But the expedition sailed under a cloud. One night, shortly before it was due to depart, most of the figures of Hermes in the city were defaced. The outrage was widely regarded as a sinister omen, and rewards were offered to informers who brought evidence about the Hermae or any other acts of sacrilege. Nothing was immediately discovered about the Hermae, but reports were brought about other offenses, including parodies of the Eleusinian Mysteries in private houses, in which Alcibiades was implicated. This opportunity was eagerly seized by Androcles and other political enemies of Alcibiades, who roused popular feeling against him. Alcibiades urged that his case should be tried at once, but his enemies, who were afraid that he would have the support of the expeditionary force, insisted that the case should be postponed until further evidence was collected. Soon after operations had begun in Sicily Alcibiades was summoned home. Realizing that the scales were weighted against him, he gave his escort the slip and soon was in Sparta putting his knowledge and advice at its disposal.

Defacement of the Hermae and the defection of Alcibiades

It was this action against Alcibiades that was largely responsible for the Athenian disaster in Sicily. Morale was badly shaken by the recall of the driving force behind the expedition, and not long afterward Lamachus was killed in battle. Nicias, whose indecisiveness was made worse by his poor health, was left in control. Progress had at first been promising, although too slow. The Athenian fleet controlled the sea, and a wall around the city was begun in order to complete the blockade. But Gylippus, a vigorous Spartan general, brought new confidence to the defenders, and Corinth also sent help. The circumvallation wall was cut off and the Athenians fell back on the defensive. Nicias, frightened to return home unsuccessful, sent for reinforcements, which arrived under Demosthenes. He made a bold night attack on Epipolae, the rocky plateau overlooking Syracuse; but his units became confused in the dark, and the attack had to be called off. Demosthenes was now anxious to leave; but Nicias hesitated, and an eclipse of the moon persuaded him to wait further. A great naval battle in the harbour was lost, and the only hope was to lead the forces overland to friendly country. But the roads were blocked, and after grim harrassment the survivors of both divisions surrendered. Nicias and Demosthenes were put to death and the men were imprisoned in quarries from which some were released for reciting verses from Euripides; others were ransomed, but many died.

Consequence of the defeat in Sicily

The almost complete destruction of the fleet and army that went to Sicily was a crippling blow to Athens. Few ships and little money were left. Urged on by Alcibiades, Agis, the Spartan king, had brought a force into Attica and was encamped at Decelea, only 12 miles from Athens itself. As a result Attica had to be completely abandoned; the silver mines at Laurium could not be worked, and some 20,000 slaves took refuge with the Spartans. Widespread revolt could be expected among the allies. The survival of Athens for eight further years is a great tribute to its courage and determination, but the Spartans, as so often before, were slow to press home their advantage. The Athenians were still safe within their walls provided that they could import enough food; this the Spartans could not prevent until they achieved decisive superiority at sea.

In order to preserve what they could of their empire, the Athenians made Samos their chief naval base, and most of the fighting shifted to the east side of the Aegean. The Spartans soon found that their problems were more complex than they had anticipated. While they were able by 412 to build up, with the help of their allies, a fleet sufficiently large to match that of the Athenians, they lacked the financial resources to maintain their crews and had to appeal to the Persians for help. This, however, posed an embarrassing dilemma. The Persian price was naturally the recovery of the Greek cities that had once belonged to the Persian Empire, and Sparta had claimed to be the liberator of the Greeks. It is not surprising that the first agreement with Persia, which recognized the Persian king's right to all that he had formerly ruled, was modified twice within six months, and that Persian pay became very irregular. Spartan relations with Persia also made the east Greeks less anxious to revolt, and Athens still had considerable support from democratic parties that knew that both Sparta and Persia favoured oligarchy.

Oligarchy in Athens. Sparta lost its best chance of an early victory in 411 when Athens was torn by political upheavals. Before the Syracusan disaster there had been no serious threat to the democracy, but conditions were now greatly changed. Periclean democracy depended on state pay for state service, and there would soon be no money to pay. The need for some restraint on the democracy had been recognized in 413 when the Assembly agreed to the appointment of ten elderly commissioners to ensure economies and guide policy. Toward the end of 412 there was open talk of the need for more substantial reform. In the fleet at Samos some of the generals and officers began to plot and approached Alcibiades, who had outstayed his welcome at Sparta and was now exercising his charm on Tissaphernes, the Persian satrap at Sardis. Alcibiades knew that Androcles and his other enemies were irreconcilable, but his family background assured him some support from his own social class. His main hope, however, rested on his assurance that he could persuade Tissaphernes to support the Athenians financially, if they changed their form of government. The oligarch Pisander led representatives of the fleet to put the case for reform before the people, and the prospect of Persian money proved decisive. The Assembly agreed in principle to constitutional changes; Pisander was sent back to complete negotiations with Alcibiades.

Cooperation of the oligarchs with Persia

When it came to the test Alcibiades could not secure the flow of Persian money, but the oligarchs had gone too far to draw back. Pisander returned to Athens, where a revolutionary situation was developing. There was much talk of a government of the 5,000 best qualified in purse and person to serve the state, but in the streets the most determined enemies of oligarchy, including Androcles, were assassinated. Pisander carried a motion that ten commissioners should be appointed to submit proposals to the Assembly on a specified day. The meeting was called at Colonus outside the walls, where, owing to the Spartan presence at Decelea, the common people would feel insecure. The commissioners merely recommended that there should be no state pay for state services, and that the normal safeguard against unconstitutional proposals should be waived. At this point, Pisander, by prearrangement, proposed that a council of 400, appointed by a combination of nomination and co-optation, should have full powers for the duration of the war with authority to convene the Five Thousand for consultation whenever they thought fit. There was no opposition, but, since delay might be dangerous, the Four Hundred entered the council house before the day appointed, with followers prepared to use force, and dismissed the democratic Five Hundred after giving them their salary for the remainder of the year.

The seizure of power was cleverly manipulated by Antiphon, an intellectual who had kept out of the public eye

in politics but had helped oligarchs in the courts and in the pursuit of office. The revolution, however, had inherent weaknesses. The voting power behind it was provided by moderates who had no wish to confine real power to a small minority; the government of the Four Hundred was created by and for extreme oligarchs who would have wished to overthrow the democracy even if the treasury was full. They had no intention of sharing power with the Five Thousand. But the Four Hundred could only remain in power if they had the sympathy of the fleet at Samos, on which Athens depended for the protection of its grain ships and the collection of tribute.

They looked to the enemy for support and planned to admit a Spartan fleet into the Piraeus, but this proved the signal for a counter-revolution led by Theramenes, who was to become the recognized leader of the moderates. The Spartan fleet sailed on to Euboea, and the small Athenian squadron guarding the straits between the mainland and the island was overwhelmed; Euboea, which had been in touch with the Spartans, now openly revolted. This was a catastrophic disaster, for a significant part of Athenian food supplies came from the island; it meant the end of the Four Hundred, whose government was dissolved by a special meeting of the Assembly. In further meetings a moderate constitution was built up, which Thucydides considered the best form of government he had known, although it lasted less than a year.

Revolt of the Euboeans

Meanwhile, the fleet from Samos had won a surprising victory at Cynossema, in the Hellespont. This success was largely due to Alcibiades, who had been recalled by the fleet. After dividing forces over the winter, the united Athenian fleet sailed into the Hellespont and surprised the Spartan fleet outside Cyzicus. The victory was complete, Sparta was swept from the seas, and the Athenians proceeded to recover the cities of the Hellespont and Propontis, over which they had lost control. To raise money quickly a station was established at Chrysopolis near the mouth of the Euxine, which exacted a toll of 10 percent on ships entering and leaving the Euxine. A further consequence was a restoration at Athens of the full democracy in the summer of 410.

By 407 Alcibiades felt sufficiently confident to return to Athens, where his reinstatement had been formally decreed. His first conspicuous gesture was to lead by land the procession to the celebration of the sacred mysteries at Eleusis, which in previous years had gone by sea in fear of the Spartan force at Decelea. He was now given supreme command with full authority to carry on the war against Sparta. Less than a year later a subordinate whom he had left in command of the fleet at Notium, with instructions not to give battle, was drawn into an engagement in which 15 ships were lost. His enemies at home seized the opportunity to stir up feeling against him, and the Assembly, which was becoming increasingly unstable as nerves became more strained, elected a new board of generals. Alcibiades retired to a castle in the Chersonese and from there watched the ruin of Athens.

Lysander and the victory of Sparta. The Battle of Notium (406 BC) had been won by Lysander, the ablest Spartan commander of the war. Intelligent, incisive, and ruthless, he succeeded, where his predecessors had conspicuously failed, in assuring adequate financial support from Persia. He won the confidence of Cyrus, the Persian king's younger brother, who was sent down to supersede the two western satraps, and with his generous help was able to restore the strength and spirit of the Spartan fleet. The Spartan naval command lasted only a year, however, and Lysander's successor was a more typical Spartan. Resenting the humiliation of dependence on Persia, he broke off relations with Cyrus. In a major sea battle off the Arginusae islands, with some 200 ships on each side, he lost his life, and his fleet was crushingly defeated. But 12 Athenian triremes went down with most of their men, and it was thought at Athens that more should have been done to save them. A stormy sea may have been the decisive factor, but scapegoats were needed, and Theramenes and Thrasybulus, who commanded triremes in the battle and had been sent to save the men, helped to steer the blame onto the generals. The Assembly, in a fit of hysteria, insisted that instead of having the individual trials that the law required, they should all be tried together. All six who were present were condemned to death.

The mishandling of this affair aroused bitter feelings at Athens and undermined the confidence of the fleet at Samos; the new generals elected for the next year commanded little respect. The Spartans had meanwhile sent Lysander back to the fleet, nominally as a subordinate but actually to command. He moved up to Abydos in the Hellespont to threaten Athens' lifeline from the Euxine; the Athenians had to follow. They encamped near the mouth of the Goat River opposite Abydos, but after failing to bring the Spartans to battle they grew careless, and Lysander was able to make a surprise attack when most of the crews were dispersed on land. All of the Athenian ships but nine were lost. Sparta now closed the route of the grain ships and the fall of Athens was inevitable. Of the Athenian allies only Samos remained loyal and refused to surrender. A generous decree was passed by the Athenian Assembly offering Athenian citizenship to all Samians, but both cities were doomed. When Athens was near starvation Theramenes was sent to negotiate with Sparta. The terms he brought back were humiliating, but Sparta had resisted the demand of Thebes and Corinth that Athens should be destroyed. The long walls were to be pulled down; all triremes except 12 were to be surrendered; Athens was to follow the same foreign policy as Sparta; and a commission of 30 was to be appointed, of whom ten were to be nominated by Lysander. This commission was to work out the details of a more oligarchic form of government that would be acceptable to Sparta.

Athens besieged: negotiations for peace

The oligarchs at Athens now had a second opportunity, but once again there was a divergence of aim between extremists and moderates, and the extremists, led by Critias, dictated events. The lesson they thought they had learned from 411 was that they must be more ruthless with the opposition, and they extended their murders from political opponents to rich citizens whose money they needed. Realizing their insecurity, they appealed to Sparta to send a garrison, but this increased their unpopularity. Meanwhile, some democrats had found refuge in Boeotia, where the leading politicians were now hostile to Sparta, and a small force under Thrasybulus occupied the border fort of Phyle as a rallying point and from there made a surprise attack on the Piraeus, which as a trading centre was firmly democratic. The population of the Piraeus welcomed them, and they successfully resisted an attack from the city; but what seemed to be the beginning of a civil war was ended by the mediation of the Spartan king Pausanias, whose primary concern was to prevent Lysander from taking control. In the settlement that he encouraged, the leading oligarchs were allowed to withdraw to Eleusis, and the Assembly resolved not to be vindictive but to bury the past. Their bitter experience under the Thirty had ensured that there would be no internal challenge again to democracy for a long time; but the Athenians had to resign themselves to the loss of their empire and to rely once again on their own resources.

The cultural life of Athens during the war. The continuing tensions of the Peloponnesian War hastened changes in the intellectual climate of Athens, the beginnings of which could be discerned in the 430s. This was the generation of the Sophists, who disturbed longstanding traditions and challenged their contemporaries to extend the rationalism that already had been applied to the physical world to man and the state. The sophists were primarily teachers who taught for pay. They professed to teach the art of politics, and the study of rhetoric was one of their main tools. The development of democracy had increased the importance of public speaking in the Assembly and in the law courts and encouraged the study of rhetoric, which included the right use of words, the emotional rhythm of a speech, and the presentation of arguments. The Sophists knew, as Aristophanes stated in a clever parody in his *Clouds,* how to make the wrong seem right and the right, wrong. Some of the Sophists were little more than fluent salesmen, but others encouraged serious thinking. By raising such basic questions as the difference between what derives from nature and what is man-made convention,

they questioned traditional morality and undermined faith in established religion; but they paved the way for the constructive philosophers of the 4th century.

Influence of Socrates

The general influences that produced the Sophists also produced Socrates, who was to have more influence on succeeding generations than any of his Greek contemporaries. Unlike the Sophists, who professed to equip their pupils for the practical problems of politics, Socrates, who was himself a poor man, did not charge fees and was concerned to make men work out for themselves true values and to consider ends before means; his method was to move toward the truth by rigorous cross-examination. Though despising current political practice, he regarded state service as a citizen's duty and himself served in the Boule; but, true to his character, when the Assembly, against existing law, insisted on trying six generals together after the Battle of Arginusae, he refused to put the proposal forward. He attracted some of the ablest young Athenians of the day, including Critias, Alcibiades, and his cousin Charmides. The cruelties of Critias when he directed the oppressive policies of the Thirty and the selfish ambitions of Alcibiades must have been a bitter disappointment to him—their association with him was not forgotten when they were dead and he was brought to trial. Thucydides, the historian, though very different from Socrates in character and interests, reflects with him the intellectual boldness of this generation. His passion for antithesis shows his keen interest in the rhetoric that the Sophists were teaching, but his mind was tougher and his thinking more profound than that of the Sophists. His *History of the Peloponnesian War* is a penetrating analysis of human nature and political motivation. He was particularly fascinated by the study of power, and from the outset in his summary of early history he emphasizes that it is inevitable that the strong should control the weak; and this assumption pervades his attitude to the Athenian Empire. In interstate relations expediency alone governs policies. He discarded the gods from his history, and for him oracles were only for the superstitious. Though less than a full generation separates him from Herodotus, he thus belongs to a different world.

Thucydides attached considerable importance to accuracy in observation and recording. In this he shared the mental attitude of Hippocrates and the medical writers who followed him. They too had removed the religious framework from their calling and looked to physical causes for physical symptoms. Epilepsy was not a "sacred disease" but a maladjustment of the brain. They built their medicine on a detailed study of cases.

Euripidean tragedy

Euripides, the third of the great Athenian tragedians, was a contemporary of Thucydides and members of the new school of medicine, and he had much in common with them. Like them, he could not be satisfied with traditional religion, but whereas Thucydides quietly ignored the gods, Euripides argued with them; the considered piety of Sophocles was replaced by a restless agnosticism. Euripides was profoundly interested in the roots of human behaviour, and in such characters as Medea and Phaedra he provided psychological studies of considerable power. But Aristotle was not being merely old-fashioned when he complained that tragedy lost something when heroic characters were modernized. Euripides also modified the form of tragedy. His plays tended to be less organic, and the chorus, whose importance was already declining, sometimes served simply as a musical interlude. The same loosening of structure can be seen in his use of the dithyramb, which adopted a freer form and developed the lyric solo, thus giving more play to the emotions.

Similar tendencies toward realism and individualism can be seen in contemporary sculpture, which tended to react against the disciplined restraint of the full Classical style and became increasingly concerned with the decorative treatment of drapery and a wider range of subjects. Of architecture too little remains to judge, but the emphasis on exquisite decoration in the Erechtheum is significant; it would have been out of place in Periclean temples. It is also no coincidence that the most famous painters of the late 5th century, Parrhasius and Zeuxis, carried realism to extremes.

Rationalism, realism, and individualism were the keynotes of this generation, but the attacks on the free-thinkers Anaxagoras, Protagoras, Diagoras, and Socrates are a reminder that conservative religious forces were still strong. Aristophanes made great fun, not always good-humoured, of the academic cleverness of the Sophists in his *Clouds* and *Wasps* and looked back wistfully on the simple instinctive virtues of the men of Marathon. But he knew in his heart that the men of Marathon would have been dull company, and he delighted in his own cleverness, which was a match for any Sophist; he would have liked to have the best of both worlds. The new attitudes weakened the close bonds of the city-state, but they were to have a fertilizing influence far beyond it. In Abdera, meanwhile, which had been the home of Protagoras (the greatest of the early Sophists), Leucippus and Democritus laid the foundations for an atomic theory that owed more to the Ionian physicists of the 6th century than to the Sophists. It is astounding that while Greece was torn by the most destructive war in its history, it should have produced such a wealth of creative achievement. (Ru.M.)

THE GREEK WORLD FROM 404 TO 323 BC

The Spartan hegemony and decline. *Spartan policy toward the Greek states.* After 27 years the Peloponnesian War had wound to a close. Sparta had overcome Athens' maritime empire and had destroyed the Delian League. The decisive battle had been fought at Aegospotami, on the European side of the Hellespont, and the defeated Athenians saw their naval and military power melt away. Blockaded and starving within their walls, they sent the wily negotiator Theramenes to Sparta to sue for peace. Peace had been rejected several times during the war by each side, and the so-called Peace of Nicias (421) had barely lasted a few deceptive years. But finally, in the spring of 404, peace was signed by the warring states of Greece at a moment when the vast Persian Empire began to be troubled by the death of Darius II, the man who had finally tipped the balance of the war in favour of Sparta. In 404 the Greek world, divided as never before, despite the peace, entered a new stage of its history.

The peace of 404

Would Sparta, victorious but weakened, succeed where Athens had failed? Sparta was a town without walls whose defense lay in the strength, moral and physical, of its soldiers. Its salient characteristics were toughness, tenacity, and insularity. The city was ruled by a narrowminded oligarchy that totally lacked imagination. But the Athenians, with all their boast of intelligence and creative energy, had shown themselves incapable of pursuing a long-term policy. Fortunately for the Athenians, they had an oligarchic party traditionally devoted to a Spartan alliance, while Sparta itself was divided into two factions: that of Lysander, victor of Aegospotami, the implacable war leader, obstinately hostile to Athens, and that of Pausanias, the king, perhaps jealous of his rival's presitige but also wise or generous enough to realize that the total destruction of Athens would create a dangerous imbalance of power in Greece.

Sparta's allies, especially Corinth and Thebes, wanted Athens annihilated once and for all. The Spartans would not go so far, but their terms were nevertheless severe. Athens kept Attica and Salamis but had to surrender what was left of its fleet and demolish the fortifications of the Piraeus, the city's port. Further, the city was to raze the long walls that defended it, recall the exiles, abandon all foreign possessions, and conclude an offensive and defensive alliance with Sparta. Once the head of a maritime empire, Athens was henceforth merely a member of the Peloponnesian League under the leadership of Sparta.

As soon as the peace was signed, Lysander made a triumphal entry into the Piraeus. He ordered the immediate demolition of the walls and called upon the pro-Spartan party to take over the government of Athens. In September 404, backed by a vote of the terrorized Assembly, he put in power a group of thirty Athenians, which included the peace negotiator Theramenes and Critias, Socrates' former pupil and first cousin of Plato's mother.

The Thirty and their regime were destined to leave behind unpleasant memories. They reduced the citizen body

to 3,000 privileged persons and for the rest behaved as tyrants. Lysander had chosen the right men to complete the dissolution of the city he hated, but Sparta sent him on a new campaign overseas, and the Thirty held power only for eight months. They reigned through terror, and their rule was one long succession of arrests, political trials, and profitable confiscations. Fifteen hundred citizens were put to death, guilty only of political opposition, and pressure was put upon the Persian satrap Pharnabazus to have the exiled Athenian leader Alcibiades murdered.

Complicity in crime makes men stand together—for a time. But, while dissension was growing among the Thirty, while Theramenes was falling out with Critias, civil war was on the point of breaking out in Attica and being loosed upon Athens. Opposition to the Thirty was led by the exile Thrasybulus, chief of the democratic party at Athens. Surreptitiously helped by Thebes, his little army grew day to day, and in May 403 he won a decisive victory over the Thirty in the Piraeus, Critias perishing in the battle.

Deposition of the Thirty

The surviving tyrants were deposed and took refuge at Eleusis, where two years later they were massacred in an ambush. Lysander reappeared and for a moment was ready to blockade the Piraeus, until Pausanias hurried from Sparta to reestablish order. After defeating Thrasybulus' forces, Pausanias was wise enough to set himself up as arbitrator between the two parties: that of the city, as it was called, the more or less faithful followers of the Thirty, which had Athens as its seat of government; and that of the Piraeus, the democrats united around Thrasybulus. Pausanias managed to bring the two parties together, and in autumn 403, under his auspices, the Athenians voted for a general amnesty, from which only the Thirty and the most heavily compromised of their officials were excluded. The old regime was restored, and in the following year (403–402) the Athenians undertook a general revision of the laws and a reorganization of the citizen body, which had suffered much from the two wars, external and civil. In 401–400 the amnesty was confirmed. Socrates, however, did not benefit from it, and he in his turn drank hemlock in 399.

Athens had thus lost its empire and its leadership of the Greek world. But, if the Greek cities thought that the time had come for them to recover or obtain their freedom, they wee soon disabused. Sparta, short both of ideas and of manpower, impoverished by a long war, and haunted by the spectre of a helot revolt, proved incapable of fashioning a new system and merely imposed its own empire in place of that of Athens.

In Thessaly the Spartans put down a civil war; in the Peloponnese they strengthened their own domination. In particular, they invaded and destroyed Elis, leaving that city nothing but its traditional responsibility for the Olympic Games. Sparta, however, was directly threatened by trouble at home. In 397 a plot led to a massacre of citizens. Supported by Lysander, who was used to plots, the new king, Agesilaus II, destined to become the great man of Sparta and friend of Xenophon, drowned the conspiracy in blood, and Sparta was free to complete the task of establishing its own domination over the Greeks. Lysander made them realize that they had merely changed their master.

It was Lysander who, since the peace of 404, had been organizing a Spartan empire through blood and war methods. He brought under Spartan hegemony nearly every Aegean government that had once been controlled by Athens. Athenian cleruchs (colonies placed in conquered territory) were replaced by Spartan garrisons, Athenian *episkopoi* by harsh governors known as "harmosts." The islands, including the sacred island of Delos, were reduced to servitude, and in the cities that had formerly belonged to the Delian League he installed oligarchic governments loyal to Sparta; ten magistrates were invested with supreme power and instructed to exact the same tribute the islands had paid to Athens. In Greece the new system was merely a copy of the old, under a different master. In Persia, however, there lived a king, not a Greek, who might have something to say about Greek affairs.

Relations with Persia. The services that the younger Cyrus had rendered to Lysander from 408 onward had not engendered any gratitude toward the Great King, and when Darius II died, in 404, the victorious Spartans no longer needed either his financial or his military support. The treaties between Sparta and the Persian king, concluded without good faith on either side, had put the Ionian Greeks at the mercy of Persia. Now Sparta, wishing to consolidate its new empire, wanted to rally under its banner all Greeks, including those of Asia. On this point the Spartans risked the hostility of Persia.

Caught between two obligations, Sparta proceeded cautiously, profiting from Persian dynastic quarrels over the succession to the throne. The younger Cyrus had mounted an expedition in Lydia to depose his brother, the new king of Persia, Artaxerxes II. By extravagant promises he had persuaded 10,000 Greek mercenaries to go with him across the deserts, first against the satrap Tissaphernes but ultimately aiming to reach the capital city, Susa, in the heart of the empire. Sparta helped Cyrus in his attempt (401), sending a small fleet and a few hundred hoplites. It was a ridiculous escapade, though it allowed Xenophon to distinguish himself by leading the survivors back to Ionia after two years of hardships. The real end of the enterprise came only eight months after the expedition set out, at the Battle of Cunaxa, where, although the King's army was defeated, Cyrus himself fell, thus depriving the expedition of any point it may have had. In any case, a local wound had no effect on the huge body of the Persian Empire. The Ten Thousand had shown, however, that with determination Greeks could venture far from the sea, cross unknown lands, and challenge the Great King with impunity. The lesson would not be forgotten.

The Ten Thousand

Nevertheless, the Greek cities of Ionia had asked Sparta, in which they saw the champion of their liberties, to defend them against the satrap Tissaphernes. The return of what remained of the Ten Thousand allowed Sparta to economize on its own citizens and to send only a small body of troops into Asia, with an inexperienced general, Thibron, who soon had to be replaced by Dercyllidas. A considerable number of Greek cities were thus freed from Persian control. The King, however, finally tired of these provocations and prepared for full-scale war. In spring 396 Sparta had to send a large army, commanded by Agesilaus. He was a brilliant tactician, and Xenophon, who served under him, invests him with every virtue both as a man and as a general. He created a body of cavalry, organized fleet and infantry, and in a two-year campaign in Ionia went from success to success at the expense of the satraps Tissaphernes and, later, Tithraustes. But Susa was still far away, across the desert, and the empire could not be shaken. Though too big to be easily defended, the empire was also too big to be conquered.

Agesilaus, moreover, was obliged to retrace his steps, furious at heart, and had to abandon any attempt to follow up his victories: ten years after Aegospotami, in 395, Sparta was again threatened in Greece itself and recalled the King to meet the new emergency.

The Corinthian War (395–387). The peace of 404, far from satisfying Thebes, had merely intensified its bitterness against its Spartan ally. Discontented, Thebes naturally sought a rapprochement with Athens, which could now be of assistance in upsetting the unbearable domination of Sparta. While Persian gold once more flowed into Greece to stir up the Greeks against Sparta, the old enemies Thebes and Athens entered into an alliance and launched a campaign in central Greece that cost Sparta the life of Lysander, drove Pausanias into exile (395), and brought into being against Sparta a coalition of Argos, Corinth, Thebes, and Athens, whose interests for the moment coincided.

Alliance of Thebes and Athens

Agesilaus, hastening back from Asia, won a brilliant victory over the forces of the coalition at Coronea in Boeotia in 394. The victory, however, was not followed up, because, at the same time, the Athenian admiral Conon, the defeated commander at Aegospotami who had since entered the King's service, crushed Sparta's fleet at Cnidus, in Ionia. On the strength of his victory, Conon returned in triumph to Athens, where he set about rebuilding the walls in defiance of the peace of 404. Thus, both in Greece

and in Asia Minor, Sparta's empire suffered severe blows at the hands of the Greeks, while that of Athens seemed capable of rising from its ashes.

To save its empire, Sparta took the easy way out. Without further hesitation, it changed policy once again and sought the help of the barbarian against the Greeks.

The King's Peace (386). Once again the internal conflicts among the Greek cities made the Great King the arbiter of Greece. There was a rush of embassies to Sardis, on the banks of the Pactolus, capital of the southern Ionian satrapy—embassies from Sparta, on the one side, and from Athens, Thebes, and Argos, on the other. To stop the Athenians, who dreamed of reestablishing their empire, the Spartan envoys supported the autonomy of the Greek cities but at the same time proposed that the Greek cities of Asia Minor come under the Great King's authority. Thus, in 392 they proclaimed their pro-Persian policy. The following year another congress was held, this time at Sparta, and again the Spartans proposed autonomy for all the Greeks. This proposal was deceptive and was not carried out, since it aimed simply at separating Athens from its new allies. The Corinthian War flared up again, and hostilities began again both in Greece and in Asia. Athens, however, was not taken unawares; under the influence of Iphicrates the city restored its fleet and modernized its army.

While Agesilaus was occupied with a hard struggle against the Argives and Corinthians in Acarnania, Iphicrates went from success to success. Another Athenian, Chabrias, was sent in 390 to Cyprus to support the petty king Evagoras against the Great King. This interference was unwise at a time when money was scarce and the courts were working full-time to make citizens pay their war tax. It presumed upon Athens' strength and at the same time pushed the Great King more toward the side of Sparta, which, in addition, received help from the west, that of the elder Dionysius, tyrant of Syracuse.

Sparta, now confident of success, sent Antalcidas on a second embassy to Sardis, then to Susa (387). The result was the King's Peace (or the Peace of Antalcidas; spring 386). It bore the stamp of Sparta and Persia, which had agreed to divide up the world. Representatives from Athens, Argos, Corinth, Thebes, and, for form's sake, Sparta were summoned to Sardis to hear the terms of the peace that the King in Susa was pleased to grant. The west was to belong to Dionysius of Syracuse; the east, with the Ionian Greeks, betrayed by Sparta, to Artaxerxes; and Greece itself passed once again under Spartan hegemony. Finally, if any city refused to sign, the King would make war on it.

Aftermath of the King's Peace; Sparta's decline. Thus, the power of the Achaemenids was imposed on all the Greek cities, both in Europe and in Asia, and even on Sparta, whose complicity merely concealed its subservience. This proud city could maintain its precarious hegemony over the Greeks only by showing itself pliable enough to deserve the Persian king's support.

But, if Sparta was pliable toward Persia, it showed a new rigidity toward the Greeks. Doubtless the King had laid down that all Greek cities should be "autonomous." But if Sparta could accept this clause, which really meant autonomy for itself and for no one else, the other cities could not accept a system that stopped them banding together against Spartan tyranny.

Spartan lack of leadership

Sparta itself, if only it had been able to escape from its ossified principles and take a clear and balanced view of the realities of world politics, would have realized that its own mistakes, aggravated by its own deficiencies, made the King's Peace unworkable. Sparta was perpetually short of men, of citizens, even of generals. Agesilaus was still on hand, no doubt, wherever his services were needed, but the hidebound system provoked inevitable conflicts between the ephors and the kings. Sparta, the military city *par excellence,* had produced no general capable of modernizing its methods of warfare, as Iphicrates had done at Athens. The obscure figures who directed Sparta's destiny could imagine no way save that of brute force to keep order in Greece. They did not realize that the Athenians were silently working to rebuild their forces. Time itself was working against Sparta, blinded by its deceptive and doomed hegemony.

The Greek cities of Asia Minor had been under Persian rule for a half century—indifferent to liberty and slumbering in commercial prosperity. The King intervened in Egypt to put down disturbances and in Cyprus to reduce Evagoras and thus deprive Athens of its last supporter in the east. Meanwhile, Agesilaus, in Greece, applied the most inflexible principles of Lysander. He laid an iron hand on Corinth, Mantineia, and Megara, and suppressed any movement for independence outside the Peloponnesian League. Conversely, autonomy was granted to the cities of Boeotia, to undermine the supremacy of Thebes.

Sparta did not realize that its successes in fact masked serious defects. Passing through Boeotia on his way north to conduct operations in Chalcidice, the Spartiate Phoebidas, secretly backed by the Spartan government, seized the Cadmea, the citadel of Thebes. This outrage, which caused a sensation in Greece, was followed by a revolution in Thebes, where Sparta took the measures necessary to ensure that no help came from Athens. The presence of Peloponnesian armies in Chalcidice was equally directed against Athens, which had to be prevented from helping the Chalcidian League. Ever since the time of Brasidas, Sparta knew how important it was to Athens to maintain its presence in these Thraceward regions. The Spartan generals sent north at this time met many difficulties, but in the summer of 379 they secured the surrender of Olynthus, a city destined to play a large role in the future history of Athens and of Macedonia.

Sparta had never appeared more powerful, but it was only in appearance. In Xenophon's words, the Spartans "could at this time believe that their complete domination was gloriously and solidly established," but, he adds, "the gods do not forget men who violate both divine and human laws. . . ; those who had sworn to leave cities autonomous were about to suffer their first punishment." Excess in fact brings nemesis.

Theban revival

In December 379 Thebes awoke with a start. Two Thebans, Epaminondas and Pelopidas, organized a conspiracy against the Spartan garrison of the Cadmea. The Spartans, taken by surprise, had to surrender. An army promptly came up from Sparta, but it, too, was defeated in front of Thebes. One of its generals decided to imitate Phoebidas' surprise attack by risking an attack on the Piraeus, which he almost captured. Greece was scandalized by this second attempt contrary to international law, and the acquittal of Sphodrias when he was put on trial at Sparta opened the eyes of all those who had blinded themselves to Spartan intentions.

Animated by a common hatred of Sparta, Athens and Thebes again concluded a defensive alliance (378). Agesilaus appeared before Thebes with a powerful army, but his failure merely strengthened the tie between Athens and Thebes that marked the end of the Spartan hegemony. It had taken Sparta a quarter of a century to bring down Athens. The same length of time had now proved sufficient for Aegospotami to be wiped out.

The year 378 marked the failure of Sparta's policy in Persia as well. Although individual Greeks had long advocated a rational crusade against the Persians, Artaxerxes was beginning to doubt the wisdom of supporting Sparta. There was little point in interfering in Greek affairs. The Greeks hated one another and could be relied upon to destroy themselves. Their threat to Persia was negligible.

The Second Athenian League; the restoration of democracy (378–371). The Athenian democracy, destroyed and then restored in 404, had never let the flame of its traditional imperialism die down completely. Helped by the satrap Pharnabazus, the Athenians had profited from Conon's victory at Cnidus (394), which destroyed Sparta's new naval supremacy, by recovering a number of islands, such as Lemnos, Imbros, and Scyros. Three years later, Andocides delivered his speech "On the Peace," in which he developed the theme of an Athens whose strength was based on the Piraeus, the walls, and the fleet, without alienating Sparta, and of prosperity through power. In 389 the theory was put into practice with Thrasybulus' campaigns in the north of the Aegean, which were hampered by the

emptiness of the treasury but which aimed to reestablish an Athenian League and thus to bring money into the state's coffers. The King's Peace in 386 had lulled Sparta's anxieties by checking this policy of patient rebuilding. But in 378 the Theban alliance with Athens revived the latter's imperial ambitions. In the intervening period they had never missed a chance of keeping open relations and, where possible, of maintaining alliances with the Aegean islands. The islands, on their side, from 386 onward saw in Athens their own defense against Persian domination. Athens had signed a pact with Chios in 384; it now formed an alliance with Byzantium, Lesbos, and Rhodes (378).

Constitution of the league

It seemed a suitable moment to find a constitutional formula that would replace these separate alliances and would encourage other new allies to join together into a single organization. Negotiations lasted from autumn of 378 through spring 377. This shows how much care went into them. Finally, the decree of Aristoteles gave the Second Athenian League its constitutional form.

The leaders of Athenian policy were basically well aware of the abuses to which Pericles' principles of imperialism had given rise, although it cannot be shown that they wished to guard against their repetition. Nevertheless, for the sake of form at least and perhaps to reassure the new members of the league, the constitution was based upon theories of federalism that were interesting by their novelty and that showed an impulse toward reform or at least a real intention of it.

The new federal organization respected the terms of the King's Peace as far as Persia was concerned but invited all peoples not under the King's rule to join the league. Their autonomy was guaranteed, putting them rigorously on an equal footing with Athens and giving Sparta no cause to interfere. No city had a foreign garrison or foreign governor. At the Synedrion, or Council, which met at Athens, each member had one vote only. On the financial side the "tribute," of bitter memory, gave place to a "contribution" to the federal treasury. These were the essential features of this first attempt in Greek history to form a representative international system with the cities sending delegates to a permanent assembly. But Athens, of course, kept its own Assembly, alongside the Synedrion, so that there existed two assemblies of equal standing in law, with that of Athens in fact exercising considerable influence. As it was, however, the organization was well-enough planned to last for 40 years, until 338, after Philip of Macedon had defeated the Greek forces at Chaeronea.

Borne up by new hopes, stirred as always by their spirit of initiative, the Athenians turned to domestic reforms affecting the Council, the law courts, the fleet, and above all—thanks to Callistratus of Aphidna—the financial system. Since there was to be no more tribute from the allies, the basis of taxation was revised, and a census of private wealth was taken. The citizens were divided into a hundred *symmories,* each of which was responsible through its own tax collectors for paying to the treasury $^1/_{100}$ of the tax levied by the state.

The lawmakers having done their job, it was time for the men of action to do theirs. Among the more prominent of them were the generals Iphicrates and Chabrias and a first-class admiral, Timotheus, son of Conon and follower of Isocrates. It was thanks to Athenians of this calibre that the new league got under way. The most extravagant hopes seemed justified, whereas Spartan disquiet grew proportionately. The Spartans had no success in operations against Thebes nor in an attempt to rebuild a fleet to regain the mastery of the Aegean. Chabrias and Timotheus, on the other hand, were winning victories everywhere, and each victory joined new cities to the league, which, young as it was, thrived to the point where it recalled the Delian League a century earlier. Nothing was missing save the Greek cities of Ionia, which the peace of 386 had abandoned to the King.

Sparta felt threatened and beneath this threat succeeded with the help of the King and of the elder Dionysius, tyrant of Syracuse, in concluding a new treaty with Athens (374): Athens recognized Sparta's supremacy in the Peloponnese, which was no great sacrifice, and Sparta, in return, recognized the Athenian League. Then Sparta changed its policy, broke the treaty, and again intervened against Athens in the west. It was merely a brush fire, however, for Timotheus, Callistratus, and Iphicrates, although bickering among themselves, soon brought Sparta to see reason.

Sparta was in full decline and no longer feared. Athens, however, had a new enemy on its hands and another in view on the horizon. In Thessaly, Jason of Pherae, thinker and man of action alike, succeeded in the difficult task of uniting the country and thus making it powerful. A still graver danger threatened in Boeotia, where at Thebes, slowly but surely, Epaminondas and Pelopidas had reorganized the army and indeed reorganized the whole country around the army, turning the old Boeotian League into a powerful and well-organized federal state. Wisdom suggested that peace in Greece should be set on a more solid basis.

Rise of Thessaly and Thebes

Sparta wanted peace, while Athens, prey to new financial troubles and worried by the growing disharmony among its best generals, was also interested. Athens, moreover, could count on the support of a league now 75 strong, though this was primarily moral support, since the members of the league did not have the financial resources to wage a war and were not prepared to sacrifice their own money once the danger from Sparta was gone.

It was thus the moment for peace. In 371 there met at Sparta a congress not merely of all the Greek states but one on a truly worldwide scale, since in addition to Dionysius of Syracuse the barbarians were also represented, notably Amyntas of Macedonia and Artaxerxes of Persia. Agreement was reached, and peace seemed assured for many years to come. But, at the last moment, one of Epaminondas' demands caused the Boeotians to be excluded from the treaty, to their great pride at this distinction. On the one side stood the world and on the other Thebes, not caring that it stood alone, because with its outstanding generals and its present strength it felt itself in a position to take over from Sparta and Athens the hegemonies they had long exercised.

Theban expansion and containment. With the stubborn patience of the peasant used to the inevitable cycle of the seasons, the Thebans had been waiting for their moment to come since 404. They had experienced the bitterness of the realization that in helping Sparta to crush Athens they had helped the most ungrateful of allies, who had then refused to share the harvest they had reaped. The King's Peace had shown them, moreover, that Sparta was capable of any treachery to consolidate its own supremacy in Greece. The peace had forced them to dissolve the Boeotian League; the seizure of the Cadmea by Phoebidas was the last straw, proving, as if further proof were needed, how little respect Sparta had for its staunchest allies. Thebes was ready for revenge as soon as the hour should strike.

Thebes could stay aloof of the peace of 371 without danger. In merely eight years Pelopidas and Epaminondas had made Thebes the most powerful city of Greece: working together, they had reorganized the army, had created the famous Sacred Band composed of 150 pairs of young men from the leading families, and had given an unusual flexibility to military tactics by inventing the charge in oblique order, by echelons, thus permitting the army to pierce the opposing line like a trireme with its ram. They also re-established the Boeotian League on a new basis and in defiance of the King's Peace.

At the beginning of July 371 the Spartan king Cleombrotus, relying on the reputation of his hoplites, unwisely invaded Boeotian territory. He supposed that with his army of 10,000 men he would have no difficulty in forcing Thebes to dissolve the league, which was intolerable to Sparta and which made a mock of treaties. But surprise carried the day. At Leuctra, west of Thebes, Epaminondas inflicted on the invading forces a blow from which Sparta never recovered. Leuctra was the first great defeat of the Spartan infantry.

Battle of Leuctra

All Greece was stupefied and flocked to join the victor. The Boeotian League was strengthened, the Athenian Confederation enlarged, the Peloponnesian League dissolved. The Spartan harmosts who still remained were driven out

or killed. Soon Sparta felt itself threatened in the very heart of the Peloponnese, as the vice slowly tightened. The Argives rebelled, Elis recovered the territories Sparta had stolen from it, and Mantineia raised its head and welcomed the advances that were made by its old enemy Tegea. At Thebes's instigation a new league of Greek cities, the Arcadian League, was planned and indeed realized. A new capital was established, Megalopolis, at the foot of Mt. Lycaeus, and a new federal constitution was introduced, followed by a police force. The first act of the new league was to call upon Epaminondas for help against Sparta.

Epaminondas came down into the Peloponnese and carried fire and slaughter to the very gates of Sparta. As Xenophon says, "The women of the city, who had never seen an enemy in arms, found the smoke of the fires lit by the enemy unbearable." Two Peloponnesian campaigns (a third was to follow) sufficed for Epaminondas to free Messenia. In 369 he again founded a capital, Messene, a new town with magnificent ramparts that made Sparta regret its pride in having no other defense than, as has been noted, the stalwart breasts and shields of its hoplites.

By climbing at the first blow into first place, Thebes upset the balance of Greece, which was always precarious. Its only difficulties were in the north, in Thessaly and Macedonia. Thessaly had revolted after the assassination of the brilliant Jason of Pherae, the ally of Boeotia, and Thebes intervened against his successor, who was allied to Athens. In Macedonia, Iphicrates seemed likely to establish Athenian supremacy. Thebes skillfully manoeuvred to secure an alliance with Macedonia, and the alliance was guaranteed by the despatch to Thebes of a hostage destined to a great future, the young prince Philip of Macedon, the future Philip II.

Everywhere else Thebes's successes were instantaneous and so widespread that they had two serious repercussions, in Asia as in Greece. Faced with the menace of Thebes, Greek envoys were sent to Susa in 367 to beg for the help of the King of Persia. Artaxerxes chose the stronger party, and by a rescript issued from Susa he opted for Thebes, against Athens and above all against Sparta, which was terrified of losing the advantages of the King's Peace. Athens, for its part, was forced to choose between Sparta and Thebes. Two years earlier, Iphicrates had helped Sparta, and Athens had granted the Spartans an alliance. In 367 the Athenians decided to stand by it, and backed it up by an understanding with the elder Dionysius of Syracuse, to assure its own security in the west.

General warfare in Greece

Despite these treaties and despite the King's intervention, the balance of power remained uncertain, and the Greeks feared a general conflagration. In 367 Thebes tried to summon a peace congress but failed; for Thebes itself would have been the arbitrator of peace. The enmity of the Greeks rekindled the war by sea and by land. By sea Athens sent a fleet to Asia under the command of Timotheus to put pressure on Persia, to obtain some alleviation of the rescript of Susa, and to support one of the satraps against the King. Epaminondas, on the other hand, built a fleet to challenge Athens' naval power and dared to launch a campaign right up to the Bosporus. Theban daring knew no bounds.

By land, war spread everywhere. The Thessalians were again restless, and Thebes had once again to intervene. The Theban victory at Cynoscephalae in 364 cost Pelopidas his life but secured peace in the north. In the south, however, the situation was becoming serious as the result of a local conflict between Elis and the Arcadians, which quickly assumed major proportions throughout the Greek world, since the sanctuary of Olympia lay in the territory of Elis. Every city took sides, and the Arcadian League itself split. This was playing Thebes's game.

In the spring of 362 Epaminondas decided on a fourth campaign in the Peloponnese, and on July 4 two Greek armies faced each other at Mantineia, north of Sparta. On the one side were contingents from Sparta, Athens, Elis, and the northern Arcadians, a total of 20,000 men. Epaminondas faced them with an equal number of Thebans, southern Arcadians, Messenians, and Argives. By skillful tactics the Theban general came within an ace of a great victory, but his death at the height of the battle left the result indecisive.

Only exhaustion stopped the war. Sparta was at the end of its tether, Athens had lost confidence in everything except the traditional supremacy of its triremes, and Thebes, with the deaths of Pelopidas and Epaminondas, had lost all chance of keeping its position as the leading state of Greece. The Boeotians could do no more than cast a regretful glance on their brilliant past as they went back to their fields. Untroubled peace confirmed a status quo of exhaustion among the Greek cities.

Peace and the balance of power in Greece (362–355). Plutarch relates that the last advice of his compatriot Epaminondas to the Thebans as he lay dying on the battlefield of Mantineia was to make peace. Peace was made but, as noted, for the simple reason that the Greeks has no strength left to make war and not because they had decided to abandon their quarrels and renounce war for good. Sparta, like Thebes in 371, kept aloof from the treaty of 362, because in its case the recurrent danger of helot revolt did not allow it to recognize the independence of Messenia. Nevertheless, while dreaming of its glorious past, Sparta did actually keep the peace.

Like Sparta and the Peloponnesian cities, Thebes turned in upon itself. Far from the battlefield, the soldier-peasant of the Boeotian League once more observed the succession of work and days. But Epaminondas and Pelopidas had left an example of how to combine force with flexibility, and this example was not lost on the young Greek-barbarian prince, Philip of Macedon, who had come to the great Thebans as a hostage and remained as their disciple and friend.

Until this young prince and, later, his son and successor Alexander had shown them the way, the Greeks in the first half of the 4th century BC conceived of no pattern of international relations except groupings of cities around one of their number that was for the moment powerful. The Athenian hegemony had ended in 404, the Spartan in 378, the Theban in 362. Moreover, none of these cities had been able to keep its pre-eminence for any considerable period except by an alliance with the next strongest. The balance of power subsisting after the Battle of Mantineia rested solely on an equal degree of exhaustion all round. Moreover, Greece had, as it happened, nothing to fear for the moment from Persia, which was disturbed by the revolt of the western provinces of its overgrown empire.

Athens, however, no doubt because its league extending overseas gave it a fleet, seemed for the moment, at least in theory, the least feeble of the Greek cities. The empire was vital to Athens' existence, but its extension created problems, and before long the Athenian League caused Athens grave worry.

Collapse of the Athenian League

The chief anxieties came from the north, from the lands far beyond Thebes, lands that always played a role in Athenian history. In Thessaly, Alexander, the successor of Jason of Pherae, was beginning to challenge Athens even by sea; but the aristocratic Thessalian League was scarcely to be feared as long as the ancient family of the Aleuads counterbalanced the tyrants of Pherae. Moreover, Athens' fleet was still more powerful than Alexander's.

Nevertheless, this fleet had need of bases around the Aegean and of free access eastward to bring the grain of the Black Sea regions to the Piraeus. Athens, however, could not succeed in regaining control of Amphipolis, now a member of another league, that of Chalcidice, under the protection of the kings of Macedonia. Another enemy of Athens was Cotys, King of Thrace, who was waging war against them in the Hellespont. Only his assassination, in 359, delivered them from the annoyance of this petty barbarian chieftain. In the direction of Magna Graecia, there were also storm clouds on the horizon. Corcyra, torn apart by its eternal factional quarrels, had seen the oligarchs get back into power and in 361 had left the Athenian League.

The Athenians in their nervousness were all too ready to turn to the old remedy of blaming the generals for the deficiencies of popular policy. The black days of the trial of the victorious generals after Arginusae seemed to have come back. The best generals were accused, arrested, and sentenced to fines, exile, or death.

Plots against the national security were often facilitated by Thebans working under cover and helped by the existence at Athens of a pro-Theban party. In 357 Thebes stirred up trouble in Euboea, the island that was Attica's natural shield and essential to the Athenian supply route. The revolt was at once suppressed, but it showed that things were sadly amiss in the league. Faced with the problem of tightening or relaxing the ties that united the allies, Athens tightened them, and they broke, resulting in the Social War, which lasted two years until 355.

In autumn 357, Chios, knowing Athens' need for money, calmly refused to pay its contribution to the league. Urged on by the ruler of Caria, Mausolus (whose tomb was later among the Seven Wonders of the World), Chios allied itself with Rhodes to the south and Byzantium to the north. So disquieting was this threefold defection that Athens sent its best generals into the Aegean. At the Battle of Embata, near Chios, the fleet failed to carry the day, which led to a trial at Athens, wherein Iphicrates was acquitted but Timotheus condemned to exile, whereupon he died. The third general, Chares, stayed in Asia to stir up dissension among the satraps and trouble the new king of Persia, who was proving difficult. But the young Artaxerxes III Ochus feared nothing; he sent Athens an ultimatum: either recall Chares or face war.

Athens, no longer in a position to choose, gave in to the demand. Thus ended the last struggle between Greeks who were supposed to be allies before they were brought into line by Macedonia. Once more, it took the Great King's intervention to bring the Social War to an end. Athens was ruined, suffering hunger and thirst. It gave up the struggle with Mausolus, recognized the independence of Byzantium, Chios, Rhodes, and of the neighbouring island of Cos, in the Aegean; of Corcyra, in the Ionian Sea; and shortly afterward of Mytilene, on the coast of Asia. All Athens kept in its possession were some tiny Cyclades, a few harbours in the north, and the right to sail freely in the Aegean. Athens had lost another empire, 50 years after the first.

The end of Athenian power

Athens had lost not only its empire but its illusions and ambitions as well. The city lived from day to day, forced to restore its finances, rebuild its economy, replace its triremes by cargo vessels. Eubulus was probably the man it needed, in 355, to recover its balance and its good sense. He was an economist and financier, scrupulously honest, a man of peace. In their weariness the people accepted him as administrator in every area of government. He struggled against fraud, waste, peculation. He balanced the two treasuries, the one reserved for the army, the other for public entertainments, somewhat to the benefit of the latter, the *theōrikon,* which helped him to maintain his popularity and thus stay in power. The improvement in the finances permitted him to undertake an extensive and useful program of public works. Athens was, however, finished as a military power.

This, however, exposed the city to danger, and a new orator, Demosthenes, saw this and began to be disturbed. For, whereas leagues and confederations had everywhere failed, the Greek world saw the rise on all sides of princes, dynasts, tyrants, and kings, both great and small, intelligent and bold, such as the elder Dionysius in Sicily, Archytas of Tarentum, Jason and Alexander in Thessaly, Evagoras in Cyprus, Cotys in Thrace, and Mausolus in Caria, not to mention Artaxerxes III in Persia. Perhaps their strength was not firmly based, but they had dared raise their heads against Athens and had on occasion shaken it.

And now appeared on the scene Philip II, he who had meditated on the example of Pelopidas and Epaminondas, overcoming dynastic problems and seating himself firmly on the throne of Macedonia, just at the moment when Eubulus' policies closed the eyes of the Athenians to the danger he presented. The year 355, like 404, opened a new era in history. It marked the end of the Athenian Empire, along with the end of the Social War. It closed an era that had seen all dreams of Greek unity brought to nought. And with the new age of Macedonia it opened new horizons.

The rise of Macedonia. *Early Macedonia.* United or otherwise, the Greek cities were a mere shadow, alongside which a new state had been gradually taking shape since the beginning of the 4th century—Macedonia, a backward country of hunters, drinkers, fighters, and peasants, whom the Greeks regarded as barbarians. The language of the country was basically Greek but so mixed with Illyrian and Thraco-Phrygian that it was incomprehensible to Athenians. Because of their violence and appeal to madness, Dionysiac rites were much celebrated. The form of government was that of a hereditary monarchy under the family of the Argeads, who enjoyed the dubious support of a great landowning aristocracy. With each generation, the kings of Macedonia tried to introduce more and more Greek civilization into their wild land.

During the Peloponnesian War, Athens had made great efforts to secure the alliance of King Perdiccas II, son of Alexander I. Macedonia could be a valuable ally. The region was rich in gold and shipbuilding timber and also lay upon the chief route for the transport of grain to several Greek cities, including Athens. The Athenians, however, might have done better to have Perdiccas as their enemy. He proved an inconstant friend and changed his coat several times before his death in 413. His son Archelaus proved, on the contrary, a great ruler, a Philhellene, and a friend of literature and the arts. He moved his capital to Pella and made it an intellectual centre to which he attracted Greek poets and musicians, such as the great tragedian Euripides, who conceived his tragedy, the *Bacchae,* at the court of this enlightened despot. Archelaus reorganized the government, centralized power, and fashioned his troops into a strong regular army. Under his reign, Macedonia ceased to be convulsed by its endemic political crises, and the historian Thucydides—a severe critic—praised him as a great builder and administrator.

Achievements of Archelaus

But, when Archelaus was assassinated in 399 by a favourite, Macedonia was plunged into 40 years of palace revolutions. Apart from its internal problems, Macedonia was also threatened from all sides: by the Paeonians to the north, the Illyrians to the west, and the Thracians to the east, while the Greek settlers in Chalcidice, to the southeast, could scarcely be considered allies; furthermore, the country lacked an outlet to the Aegean.

Amyntas III (reigned *c.* 393–370/369) had great difficulty in consolidating his throne and sought help first from Sparta, then from Athens, whose rights to Amphipolis he recognized in 371, one year before his death. His elder son and successor, Alexander II, was killed by his mother Eurydice's lover, Ptolemy, and the pair of them ruled as regents until Ptolemy, in turn, was assassinated by the second son of Amyntas III, the young Perdiccas III, then aged 20. This young king, a Philhellene and lover of philosophy, called on the exiled Athenian Callistratus to reorganize his treasury and followed the example of Archelaus in making Macedonia a strong power, until he was killed in 359 after a reign of six years in a campaign against the Illyrians, who were occupying part of his kingdom.

Problem of the succession

Macedonia was again in danger; but it can be supposed that these trials were tempering the character of the Argeads and forming a dynasty that was to produce Philip and Alexander. The problem of the frontiers was serious, but that of the succession was critical. In 359 several claimants sought the throne, supported by foreign armies. Perdiccas' son Amyntas was too young to be king, but the throne was claimed by his paternal uncles: Archelaus, Arrhidaeus, Menelaus, and Philip. Philip, the youngest, had spent three years as a hostage at Thebes, the friend of Epaminondas and Pelopidas. Philip was appointed regent for his nephew Amyntas and thus represented the legitimate power. Few could have guessed that this young Macedonian prince, aged 23, would, in less than 20 years, shape a Hellenic nation united under his own absolute authority.

The rise of Macedonia under Philip II. Unscrupulous, intelligent, determined, backed by no assembly, Philip succeeded where no purely Greek people had done so. In the past the cities of Greece had merely formed alliances or leagues under temporary hegemonies that masked or failed to mask the subjection of the weak. Philip, by imposing a new domination on the Greeks, achieved for the first time a new unity whereby all were equally enslaved.

At 23 Philip was already experienced in politics and administration, having governed one of the provinces of Macedonia after his return from Boeotia. It took him only three years to oust his nephew from the throne. When in 356 the princess Olympias gave him his hoped-for heir, Alexander, he exchanged his regency for the title of king. This gave him the green light for his hopes and desires. He could devote himsellf to his many passions, of which the foremost was the greatness of his country.

He was an exceptional character, who held all the trump cards. Educated, civilized, Hellenized, indefatigable, he was also a *bon vivant* who loved wine and women and who knew how to use his charm on strangers. He had the flexibility of the diplomat, the steadfastness of the soldier, and all the qualities of leadership. Patiently and methodically, he tackled his problems one by one. He began by re-establishing order in his enormous kingdom. (Macedonia was at that time the largest and most populous state in the Greek world.) With power centralized in his own hands, he made Macedonia even larger, at the expense of its neighbours, and Hellenized the lands he conquered. Towns that resisted him were destroyed, and new ones were founded in their stead. He favoured urbanization, built roads, and encouraged agriculture. His peasants made excellent soldiers—trained, disciplined, and full of admiration for their commander in chief.

Philip's primary need was money. He balanced his budget in the simplest and most effective way by conquering Mt. Pangaeum, whose gold mines produced 1,000 talents a year. He then founded the nearby town of Philippi in place of Crenides. He gave his name to the golden staters that came from the mines, for he now began to coin money, and the handsome "philipps" became an instrument of his policy. He understood and practiced the art of buying consciences.

Reorganization of the army

His closest care was reserved for the army, in which area he was assisted by his faithful general Parmenio. Compulsory service was introduced, each region furnishing its quota of men. Three recruiting areas were established, each one producing a unit of light infantry, one of heavy infantry, and one of cavalry. The tactics, based on those of Epaminondas, combined flexibility of manoeuvre with the thrust of light and heavy troops. The keystone was the famous Macedonian phalanx, bristling with the long lance known as a *sarissa,* which presented to the enemy an impregnable wall of iron. Once the opponent was checked and held by the phalanx, the mobile troops had merely to attack in the rear. Siege tactics also progressed. Engineering and artillery now contributed to the speedy capture of the besieged town.

At the head of this superb army, Philip II took part of Epirus and extended Macedonia's southern frontiers at the expense of Thessaly. To guard his western flank against the Illyrians, he had recourse to a diplomatic marriage with Olympias, the daughter of an Illyrian king. Philip also realized that Macedonia needed an outlet to the sea. As regent he had recognized Athens' right to Amphipolis, though with mental reservations. Now, as king and sure of his army, he attacked the town without hesitation. It was vital to Athens and had already seen much Athenian bloodshed. Philip took it, promised the Athenians to give it back in exchange for Pydna, and finished by keeping both Amphipolis and Pydna. Paralyzed by the Social War, the Athenians could do nothing but accept a *fait accompli.* Nor did they even intervene in time to save Olynthus or Potidaea. Olynthus, capital of Chalcidice and of the Chalcidian League, threw itself into Philip's arms, and he bought the Olynthians' loyalty at no cost to himself by making them a present of the territory of Potidaea, which he had had destroyed.

While Philip, almost as soon as he was on the throne, became master of a unified, powerful, and populous Macedonia, the Athenians remained blinded by the pacifist policy of Eubulus. Only the orator Demosthenes, now entering on his political career, realized the danger. He believed that Athens should keep at peace with Persia in order to concentrate all its forces against Macedonia. Only Athens could stop Philip. The Athenians, however, would not face the danger that grew clearer day by day in the north, despite the importance of this region for their grain supply.

The Third Sacred War. Philip bided his time, waiting for a chance to intervene in Greece. He found his pretext in the so-called Third Sacred War, between the Phocians of Delphi and a coalition of Thebans, Thessalians, and Locrians. Ostensibly fought to defend the rights of an outraged god, Apollo of Delphi, the war in fact was fed by an old and deep-seated hatred between Thebes and Phocis. But the religious character of the war threatened to engulf all Greece, instead of uniting the Greeks against a common enemy.

In 356 the Phocians had laid sacrilegious hands on the gold and silver in the treasuries of the Panhellenic sanctuary of Delphi, which lay in the middle of their territory. With this they made war, raising and maintaining an army of 10,000 men, which was for some time the best in Greece. Under energetic commanders this army dared invade Thessaly and Boeotia, with more successes than failures.

This attracted the wolf into the sheepfold. In 354 Philip had taken the town of Methone, at the foot of Mt. Olympus, and in Athens Demosthenes had realized the gravity of the danger from Macedonia. Philip, moving down into Thessaly, lost a first battle against the Phocian army but soon had his revenge, and all Thessaly came under his sway. Thebes, the enemy of Phocis, had naturally taken his side, Athens and Sparta that of Phocis.

The war ranged Phocis against Thebes at the front of the stage, with Macedonia versus Athens behind them. It also ranged policies and men against each other; the great duel between Demosthenes and Philip was to last almost 30 years. It was a dramatic duel, if ever there was one, but unequal, since Philip had the support of his army, his diplomacy, his undivided aims and control over his own decisions, whereas Demosthenes was armed only with his eloquence and was surrounded by other opponents, the worst of all among his own compatriots. At Athens an age of generals had given place to an age of orators, some of whom, such as Aeschines, had been bought by Philip, while others, such as Eubulus, supported by Isocrates, did not wish to see the danger and played on the people's apathy. At each step Philip made, Demosthenes tried in vain to shake the Athenian Assembly from its inertia by a speech compounded of violence and reason, but his eloquence was powerless against the strength of the King of Macedonia.

Philip and Demosthenes

The latter's intentions were clear, at least to those who were willing to see, the day when he crossed Thessaly, which he had annexed, and appeared before Thermopylae (353). This bold stroke produced a reaction in Greece, and in a reminiscence of the glorious days of the Persian Wars Philip found himself faced by the Phocian army, reinforced by 10,000 Greeks from Sparta and above all from Athens. With the way forward barred, Philip saw that the fruit was not yet ripe. He withdrew, patiently awaiting a better opportunity.

Meanwhile, he strengthened his position in the north and east, pushing into the Chersonese and the Hellespontine regions, reaching as far as Byzantium, where he opened negotiations designed to undermine Athenian influence. At the same time, he obliged the King of the Odrysians in Thrace to sign a treaty of alliance. With his hands free on that front, he could turn on a new prey, his supposed allies, the Chalcidians. He in fact needed to get the Greeks of Chalcidice into his power, not only because its promontories offered threefold access to the sea but also because its geographic situation made it an intolerable barrier between Philip's own possessions, Amphipolis, Philippi, and Pangaeum, to the east, and Pydna and Methone, to the west.

Hoping for assistance, the Chalcidians concluded a treaty with Athens. Demosthenes delivered his first "Philippic" (351), followed by the three "Olynthiacs" when Philip revealed his intentions on Olynthus, the most powerful of the Chalcidian cities. To stop Athens from interfering, Philip in Machiavellian style formented a revolt in Euboea. Athens in fact sent three expeditions to Olynthus, but it still surrendered (348). Philip razed it to the ground

for daring to resist, sold the inhabitants into slavery, destroyed 32 other Chalcidian cities, and incorporated the country into Macedonia. The Athenians saw their access to the north definitely cut off, and there was nothing to stop Macedonia replacing them as a maritime power.

Philip was then clever enough to make people believe that he would not abuse his victory. Like a good player, he sheathed the sword and turned upon Athens a friendly countenance. Many Athenians, following Aeschines, did not wish to undertake the necessary effort to keep independence and preferred to see Philip as a friend and come to terms with him. But all the rest as well, whatever their party or persuasion, equally desired an agreement with Macedonia, even Demosthenes himself, who realized how weak his country was at this time.

Negotiations were therefore opened, and the two governments played out a diplomatic comedy at Athens' expense. Athens sent an embassy to Pella, made up of Philocrates and of the two rivals in eloquence and opponents in policy, Aeschines and Demosthenes. The splendid reception that the King gave them won them over, except perhaps Demosthenes, who, however, lost his wits, either from illness or emotion, and was unable to speak. Philip's charm had done its work on the others. In return, a Macedonian delegation was sent to Athens to come to precise terms. These were debated by the Assembly, which in the confusion voted for the status quo between the two states. This meant recognizing the loss of Amphipolis and Chalcidice, while at the same time the Phocians were excluded from the treaty, which allowed Philip the latitude to continue or halt the Sacred War. But the peace was not yet signed, as the King's oath was still lacking.

Thus, a second embassy, made up of the same men as the first, set off for Pella. Demosthenes had realized that Athens must bring the affair to a swift conclusion, since it was in Philip's interest for it to drag on. He was in the process of completing his conquests in Thrace, and each day his share of the status quo grew larger. He also knew how to spend his money to cause delays. The Athenian ambassadors spent a month and a half running after him. It was only in July 346, a year after Plato's death, that Philip signed the peace known as the Peace of Philocrates. He signed in the name of the Macedonians and their allies, including the Thebans, with Athens alone as the other party. Although he was now Athens' ally, Philip could still complete the conquest of Phocis, thus becoming master of Thermopylae and of the sanctuary of Delphi, from which he expelled the Phocians permanently. His political supremacy was thus strengthened by a moral and religious authority in Greek eyes, since he now controlled the oracle of Apollo and presided over the Pythian Games of 346.

The Peace of Philocrates

Such was the result of the Sacred War. Philip's triumph, which he owed to his diplomacy and his army, ensured his complete sovereignty in northern and central Greece, and at Athens itself he increased the number of his supporters. Isocrates, the teacher of rhetoric, saw in him the man of destiny capable of uniting all the Greeks and leading them in a great crusade against the King of Persia, although the danger from Persia had ceased to exist for Athens, at least compared with that from Macedonia. Demosthenes was not deceived; but he realized that Athens, weakened as it was, had no alternative but peace, unreal though it might be.

Renewal of Athenian resistance. So Athens and Philip made peace. They even became allies. But the causes that had made the two states enemies and would make them enemies again had not been removed. Philip was still eaten up by his ambition to become supreme lord of the whole of Greece. In all Greece he found not one single city to oppose him; and Demosthenes was a man without support, save that of his own intransigent patriotism.

Isolated and obstinate, Demosthenes fought on every front. In the courts he launched a case against Aeschines over the embassy that had negotiated the Peace of Philocrates, accusing him of treason. But the citizen jury, afraid of arousing Philip's anger, acquitted Aeschines. In the Assembly Demosthenes once more denounced the Macedonian peril, courageously delivering the second "Philippic" (344). He was also active on the diplomatic front, to south and north. He sought to draw the cities of the Peloponnese into a union against Macedonia and judged it essential also to erect a living barrier between Philip and Athens. For this he had to effect a great reversal of alliances, forget the grievances against Thebes, forget that Thebes had profited from the Peace of Philocrates as ally (since 354) of Macedonia, and obtain at any cost an alliance with Thebes.

Such a reversal could not be achieved in one day, and bit by bit, while waiting, the meshes of the Macedonian net tightened. Philip negotiated with the King of Persia, stirred up fresh trouble in Euboea, and extended his threatening shadow toward the coast of Ionia and, bypassing Athens, up to the Isthmus of Corinth. Demosthenes, however, stood in the breach, persuading the cities to conclude defensive alliances with Athens. With characteristic flexibility, Philip did not insist but changed his front. Turning toward the Bosporus, he reduced Thrace to a Macedonian province (341). Next, he launched a campaign against the Thracian Chersonese, Athens' last possession in the northeast. He went in fact as far as to lay siege to Byzantium, which, thanks to Demosthenes, had rejoined the Athenian alliance. This intolerable audacity finally shook the Athenians out of their apathy, and they decided to send a fleet into the Hellespont. The Peace of Philocrates was truly at an end. Demosthenes now redoubled his activities: he delivered the last two "Philippics," brought into being a Hellenic league, improved the situation of the war chest, and restored the strength of the Athenian fleet, thanks to a reform of the trierarchy system. Byzantium was saved.

In 339 a fourth sacred war, called the Amphissean War, broke out in Greece as the result of a sacrilege committed by some men of Amphissa against the Delphic god. Apollo had to be avenged. Philip accepted the bargain. Moreover, the command of the army charged with punishing the guilty Amphisseans was conferred upon him. He secretly outflanked Thermopylae and by a surprise gained possession of Elatea, in northern Boeotia. Panic seized the Athenians: Philip was three days' march from the frontier of Attica; he had only to cross Boeotia, and his army would be in front of the walls of Athens.

The Amphissean War

In the general consternation, Demosthenes kept his head. Philip had sent to Thebes envoys loaded with promises. Demosthenes got himself sent there as well and proved to the worried Thebans that Philip's advance must be stopped. His eloquence won the day: Thebes chose Athens. Their common enemy could not tolerate such an alliance, and war broke out openly between Philip and the new allies. Fortune seemed at first to favour the Greeks during the winter campaign; the armies of Thebes and Athens won two minor victories, which had no follow-up. Instead, there ensued a fierce battle in April 338 on the plain of Chaeronea, in Boeotia, where the Macedonian phalanx routed the ill-led allied troops. It was the end of Greek liberty.

Philip's hegemony over Greece. Thebes was occupied by a Macedonian garrison. Athens got off more lightly: by what is known as the Peace of Demades, Philip, no doubt influenced by what he owed to Athenian civilization, paid homage to the conquered city and to the dignity of its resistance by granting easy conditions. He compelled Athens to sign an alliance with him and took the Chersonese; but neither the city nor Attica was occupied, the Athenians kept their fleet, and their ports remained free. Demosthenes' anti-Macedonian policy may be considered to have been justified by the leniency of the peace. If Athens had listened to Aeschines, it would have been humiliated without a fight. With Demosthenes it won the respect of a conqueror who was more Greek than barbarian. Alexander would not forget the lesson in moderation his father had given him.

After his victory Philip did not halt. Master of Athens and of central Greece, he conquered Euboea and then descended upon the Peloponnese, where he imposed his will without a fight. Greece lay at his feet.

To succeed in his program, Philip attempted to establish a permanent organization of the Greek cities under Macedonian hegemony. He summoned to Corinth, the central city, delegates from all Greece. Apart from Sparta,

all accepted the invitation to this extraordinary congress. There, in 337, there came into being a league truly "Panhellenic," since it included, apart from the Spartans, all those who henceforth called themselves officially Hellenes. Philip thus realized the ideal of Isocrates, though with one major reservation, namely that the Hellenes, though united, were united only under a foreign dominion, that of the King of Macedonia, now become their *hēgemōn* and *stratēgos* and wielding executive power. The league was represented by a council, the first example of a supranational institution or league of nations, but was entirely in Philip's hands, beneath the banner of Hellenism.

The states that adhered to the League of Corinth were declared autonomous. They paid no tribute, and their military obligations were reduced, consisting only of an obligation to provide a contingent to the federal army and fleet; it was forbidden to take up arms against Greeks, since all Greeks were henceforth linked in a general peace.

The new institutions were skillfully arranged, and Philip was diplomatic enough to cover his iron hand with a velvet glove. He was also realistic enough to realize that a common enemy was needed to strengthen the ties among men so diverse and to free the Asiatic Greeks. In 336 he proposed at Corinth a general crusade against the barbarian. A Macedonian army of 10,000 men set out at once as an advance guard under the orders of Parmenio and Attalus. Penetrating Ionia, it seized Ephesus. The Achaemenid Empire seemed to be nearing its end.

Assassination of Philip

But Philip in his universal victories had counted without fate. At the beginning of the summer of 336, when he was 48, during the magnificent celebration that he had organized for the wedding of his daughter Cleopatra to the king of Epirus, he fell beneath an assassin's dagger. Philip was dead, but Hellenism, on the contrary, had a chance to survive. It was not extinguished in the country of Isocrates and Demosthenes, and it vibrated in the heart of the young Alexander, Philip's son and heir. Alexander was 20 years old. He had been 18 at the battle of Chaeronea and 13 when Philip had appointed Aristotle his tutor. Aristotle's pupil was to Hellenize the east. But, on the other side of Greece, Hellenism still shone, with its defects and its virtues, among the western Greeks.

The western Greeks. *Conflict with Carthage.* After the Peloponnesian War, the cities of mainland Greece continued jockeying for supremacy, and the king of Persia never hesitated to intervene. It took Philip of Macedon to stop the process. The history of Greeks and barbarians in western Greece (*i.e.*, in Magna Graecia, southern Italy, and Sicily) reflected in its record of internecine warfare and barbarian intervention that of their mother cities in Greece itself. To understand the history of the 4th century, one must go back to 480, when the Carthaginians were defeated at Himera in Sicily. At the same time, the Persians were attacking the mainland Greeks. In 413 the Athenian disaster at Syracuse again produced a similar twofold barbarian intervention against the Greeks. While the King of Persia, with his satraps in Ionia, was working against both Athens and Sparta to ensure his own domination by the exhaustion of both rivals, the Carthaginians, who had been temporarily threatened by Alcibiades' plans, now on their side, breathed new life into their own Sicilian ambitions. This was a serious threat, since Sicilian unity had not survived the victory over Athens in 413.

Hermocrates

Hermocrates, the leader of the aristocratic party at Syracuse, had emerged by luck and the help of Sparta as the brilliant architect of victory. Scarcely had he turned his back to go to the Peloponnese—there to continue the struggle against Athens—when the ungrateful Syracusans exiled him and called in the democrat Diocles. On a wider scale, the same conflict also set one city of Sicily against another. Syracuse turned on Catana, Selinus on Segesta. And the Segestans called on Carthage for help against their fellow Greeks.

Carthage did not miss this chance of avenging the insult that the petty tyrant of the petty town of Gela had inflicted on it at Himera 70 years before. In 409 a Carthaginian army under Hannibal, son of Gisco, disembarked in western Sicily. It had no difficulty in beating the Selinians, whose town was promptly sacked. Himera, threatened in turn, was helped by Diocles of Syracuse but fell to superior numbers and was razed. Hermocrates, returning home, recaptured Selinus but died at Syracuse in 407. His supporters, including a young officer named Dionysius, were exiled. The following year the Carthaginians, aiming to end all resistance of the Sicilian Greeks, sent out a new expeditionary force under Hannibal and his grandnephew Himilco. They laid siege to Acragas, where Hannibal died, but Himilco's gold bought the treachery of Sicilian generals both at Syracuse and at Acragas itself, and Himilco took the town in December 406. At this moment, a saviour arose in the person of Dionysius, the young officer who had served his apprenticeship in arms with Hermocrates.

The rise of Syracuse under Dionysius. Seizing his opportunity, Dionysius, known as Dionysius the Elder, with the help of others of Hermocrates' supporters (such as the rich Philistus and Hipparinus, the father of Plato's young friend Dion), began by denouncing the treachery of the Syracusan generals. With nine new colleagues (whom he was soon to get rid of), he was elected general. He recalled the exiles, married one of Hermocrates' daughters, and organized a plebiscite. With the assistance of a group of supporters who shared his conviction that Syracuse needed a firm government based on the army, he surrounded himself with a bodyguard, by means of which in the summer of 405 he seized absolute power.

Struggle against Carthage

To retain power, he needed to remove the Carthaginian menace. Himilco had continued the conquest of Sicily, defeating Gela and Camarina. Dionysius did not rush to their help with any great haste, perhaps to leave open the possibility of an agreement with Carthage. In his absence, however, the Syracusans rebelled and killed his wife. When he returned, he put down the rebellion and succeeded in concluding an advantageous treaty with Himilco. Half of Sicily was abandoned to Carthage, but the independence of the cities in the eastern half was guaranteed and strengthened by the Syracusan tyranny, which Dionysius was to retain for 38 years, until his death, in 367.

Dionysius' first task was to establish his power on a firm basis. He reorganized the army by introducing mercenaries and fortified the royal palace on the island of Ortygia (which almost completely shuts in the Great Harbour) and fortified the Great Harbour itself. He erected defenses for the whole town by surrounding with enormous walls the extensive plateau of Epipolae, where the fortress of Euryalus became recognized as a masterpiece of military engineering. The philosopher Plato was to stay in this palace on several occasions from 388 or 387 onward.

He needed also to ensure the succession, since he proposed to turn his tyranny into a hereditary monarchy. With this in mind he turned bigamist, marrying on the same day Hipparinus' daughter Aristomache and Doris of Locri Epizephyrii, who was destined to produce the future heir Dionysius II.

If the elder Dionysius was almost wholly devoid of scruples, nevertheless it should be said that, on the whole, he used his power with moderation, once he felt that it was assured, after a final Syracusan revolt in 403.

As master of Syracuse, his ambitions grew. He wanted to rule all Sicily and make it, in Plato's phrase, "one single city." In this he merely anticipated on a smaller scale Philip's plan to make mainland Greece a single state. To this end he reduced Catana, Naxos, Leontini, and Messana and became, under the title of "archon of Sicily," the head of the citizens of the Sicilian Greek cities and also the protector of the original Sicilian inhabitants, insofar as the latter were not under Carthaginian rule. In this he anticipated Alexander as well, since he did not hesitate to deport entire communities in order to realize the fusion of peoples and races.

Projects in southern Italy

Since he still had plans for southern Italy, he proposed, without giving up the struggle against Carthage, to place some of the Italian cities under the control of Locri, home of his wife Doris; with the others, he would not deprive them of their liberty but would take them under his protection by bringing them together in a league presided over by Tarentum, under the aegis of his friend Archytas. He therefore allowed the latter, who combined the role

of politician with those of Pythagorean philosopher and mathematician of genius, to give Tarentum a mixed constitution that made it the leading city of southern Italy. Then, after concluding pacts with the Messapians, Campanians, and Celts, he founded the towns of Ancona and Adria on the Adriatic. He was anticipating not only the two great kings of Macedonia but also the Romans, since he would doubtless have brought about the unity of Italy if throughout his reign he had not had to cope with several Punic wars, as the Romans were also obliged to do later.

The treaty of 405 had not in fact finally put an end to the danger from the Carthaginians. Dionysius began by calling on Carthage in 398 to evacuate the Greek cities of Sicily, and after a siege he captured Motya, which was defended by Himilco's fleet. Himilco, however, took his revenge the following year by destroying Messana and defeating Dionysius in front of Catana, which enabled him to besiege Syracuse. But Dionysius in 396 regained the initiative, destroyed Himilco's fleet, and allowed him to escape to Carthage, where he committed suicide, unable to bear his defeat. Dionysius then took Catana, rebuilt Messana, and occupied Tauromenium (present-day Taormina), founded by Himilco five years earlier.

A second Punic war (383–376) went less well for him. He defeated the Carthaginians at Cabala, where Mago was killed; but Mago's son, the younger Himilco, got his revenge at Cronium. The elder Dionysius was forced to ask for peace and to allow a third of Sicily to pass once more under barbarian rule. A third Punic war, toward the end of his reign, proved indecisive. The status quo was maintained, and Syracuse and Carthage shared the rule over Sicily.

Relations with mainland Greece

The elder Dionysius did not, however, have to do exclusively with Italy and North Africa. The mainland Greeks, like those of Sicily itself, kept their eyes always fixed on this notable tyrant who dared to intrude upon their affairs. He remained faithful to the Syracusan alliance with Sparta, closely followed Persia's intervention in Greece at the time of the King's Peace (386), and sent a representative to the congress at Sparta (371). Athens tried to seduce him from his Spartan alliance and to get the help of his mercenaries. It crowned one of his tragedies, *The Ransom of Hector,* which he had had presented at Athens during the Lenaean Festival of 367. Although Lysias had spoken harshly of him at the Olympic Games in 388, Isocrates in 369 wrote him an open letter to invite him, as he was later to invite Philip, to head a crusade against Persia. After his death, Xenophon followed with the treatise known as the *Hieron,* ending with an encomium of the enlightened despot. Even Plato made his condemnation of tyranny less harsh on his account, and it was his realization of the impossibility of achieving anything in Athenian politics that made him undertake in 388/387 his first journey to Syracuse, where he hoped to find the seat of the philosopher-king or king-philosopher. There he met the tyrant's brother-in-law, the young and brilliant Dion, aged 21, on whom he was to found so many hopes. His journey ended badly, but on returning to Athens he founded the Academy and continued to exchange letters both with Archytas of Tarentum and with Dionysius the Elder.

This remarkable tyrant may have had excessive literary pretensions and thought himself a greater poet than he was. But his political achievements were considerable. He was a great statesman who made Syracuse the richest and most populous city in the world. His example had some influence on Alexander, since Dionysius accomplished in the west what Alexander was to repeat on a larger scale 40 years later in the east. Like Alexander, Dionysius followed the principle of bringing together Greeks and barbarians.

The western Greeks to the beginning of the 3rd century BC. In Dionysius the Younger, or Dionysius II, the elder Dionysius did not have a worthy successor. The younger man was cultured but had no morality, nor had he any talent, either political or military. He welcomed Plato, who had accepted Dion's invitation to visit Syracuse a second time, still haunted by his ideal of the philosopher-king. But Dion, whose crime was to have tried to keep his nephew the tyrant from a life of debauchery, was exiled, and Plato, once more disappointed by Syracuse, had to leave a court that was wholly given over to intrigue, in order to take up again his teaching at the Academy. Dion followed the courses, while keeping in touch with the leaders of the opposition at Syracuse.

The visits of Plato

Hoping to reconcile Dionysius and Dion, Plato risked a third journey to Syracuse, with the idea of demonstrating to the tyrant the primacy of the soul over the body and of creating with him the republic of his dreams. Once more disappointed, he hurried back to Athens (361/360). Three years later Dion recruited mercenaries and assembled a fleet to obtain a return by force. He received a triumphal welcome at Syracuse, besieged Ortygia, won a naval victory over Dionysius, and after various ups and downs had himself made general with full powers. He thereupon declared that he was applying Plato's principles but in practice seemed to forget somewhat that he was his pupil. He got rid of a dangerous subordinate but fell victim to his Athenian friend Callippus, his host at Athens, and a year after taking power at Syracuse he was assassinated by his own soldiers (June 354).

Once again anarchy reigned in Sicily. Callippus, although an Athenian, made himself tyrant of Syracuse for 13 months. He had to fight against Dion's family and their supporters under Hipparinus, who took the town and regained power until his death, two years later. Hipparinus' brother Nysaeus reigned for five years (until 347), when the younger Dionysius returned and regained the tyranny for two years. In 345 the Syracusan exiles appealed for help to their mother city, Corinth, and secured the despatch of the Corinthian envoy Timoleon, who in 344 forced the younger Dionysius' abdication. The work of the elder Dionysius was destroyed, and Sicily, following a course only too typically Greek, returned to its discords.

Timoleon

Timoleon, however, surprised everyone. This Corinthian was a great man of the type celebrated by Plutarch, perhaps painted better than he was by history, probably enamoured of the ancient virtues; once master of Syracuse, he tried to reestablish peace in the city and to restore Sicilian unity. The initial success of his program reawakened Carthaginian fears, resulting in a new Punic war. Timoleon was lucky enough to win a great victory near Segesta that compelled the Carthaginians to ask for peace (341). He could then give himself entirely to his work for peace and civilization. He freed the cities of Sicily, brought in new blood with 60,000 settlers, distributed land, and established a wise and moderate regime in Syracuse. After seven years, considering that Sicily no longer needed him, he signed his own abdication, respected by all (337).

Timoleon had restored Sicilian unity, but his influence had not spread to southern Italy, where trouble continued. Despite this, Hellenism made steady progress, gradually winning over the barbarians by spreading among them moral and intellectual ideas. Hellenism had also reached Africa, where Cyrene shone by its prosperity, its schools, and its trade, as well as Gaul, where the Provençal centres of Saint-Blaise, Olbia, Antipolis, and Agde maintained a flourishing trade with Greece and Sicily.

Before losing its independence, Greek Sicily had one last fling. While the Greco-Macedonian monarchies were taking shape in the east after Alexander's death, the freebooter Agathocles temporarily halted the decline of the island, which was doomed, with southern Italy, to annexation by the Romans. In 318/317 he seized power in Syracuse, had himself given the title of king, reopened with some success the Punic War, and was summoned by Tarentum to restore order in southern Italy. He was the man to save the Greek west and to establish a Greek "Kingdom of the Two Sicilies."

But Agathocles left it to the Romans to complete the job and to rekindle in the new capital of the world the fire of Hellenism that had burned brightly but fitfully at Syracuse.

Greek culture and society in the 4th century BC. *The changing style of life of the polis.* Philip's success in uniting the Greek cities under Macedonian domination was the result, above all, of subtle diplomacy linked to force of arms. What he had laid low, however, was merely a giant already shaken by the struggles that had scarcely stopped tearing apart the Greek peoples, separately or in

their leagues, from the Persian Wars to Philip's own time.

Since the victories never lasted, nor did the peace, the conquerors and the conquered suffered equally from the ills that war brought about: the conscription of farmers into the army, the devastation of the countryside, the flight from the land. The peasant and small landowner tended to disappear. Men returning from war were unwilling to take up the plow and instead hired themselves out as soldiers; after living dangerously but well in times of pillage, after seeing new, wider horizons, a man lost the taste for a narrow existence on land that demanded daily toil or felt himself stifled by being shut in within city walls. Mercenary soldiering was an inevitable evil if it was the means whereby alone men could live, but it contributed to the death of both country and city.

Decline of the city

The city in itself, moreover, was leading a precarious existence. Perpetual warfare slowly but surely brought economic stagnation, and the maritime leagues no longer brought in the easy tribute of the previous century. At Athens, the rich citizens were obliged to band together in order to meet the crushing charges that taxation imposed. The poorer classes expected the state to provide for them, and the level of public morality declined. The thousands of citizens who formed the juries in the lawcourts lost the habit of working, since they were paid for performing their duties as jurymen. The number of cases multiplied, while the number of informers living on blackmail and confiscations increased. Speakers had no difficulty in getting the people to vote the reestablishment of the Theoric Fund destined to pay for the people's pleasure, and this same people henceforth received a salary for being present in the Assembly. The 4th century provides the spectacle of the slow ruin of public spirit. Perhaps the life of the city-state was already moribund at the time of Philip's conquest.

The incessant wars between Greek cities and even within cities undermined by civil strife showed clearly that no regime, whether oligarchic, such as Sparta, or democratic, such as Athens, had been able to maintain equilibrium in Greece. After the death of Socrates, guilty, despite having done his duty on the battlefield, of rising above the city walls to think of man instead of defending the democracy, historians and philosophers had clearly defined the social and political problem. For the philosopher Antisthenes, founder of Cynicism, Greek and barbarian were equal, as were slave and free citizen. Among the intellectuals of the period and among those men of action who reflected upon the historical scene in which they were acting, there was born and took shape the idea that the separateness of the city-states had had its day.

Already in 392 the Sophist Gorgias took the occasion of the Olympic Games to proclaim the need to bring about a union of all Greeks, a position later taken up also by Lysias and Isocrates. The latter had at first thought the Greeks could be united under the leadership of Athens, but the history of his times gradually converted him to the idea that Greece would give up its divisions only under the direction of one single man.

The idea of Greek unification under a strong individual was also taken up by Plato, who thought for a time that he had found his philosopher-king, in Sicily, in the person of Dionysius the Elder, then of Dionysius the Younger, then of Dion. Isocrates placed his hope in Jason of Pherae, then turned his eyes on the son of the king of Sparta, Archidamus, and then again on Dionysius the Elder, who, tyrant though he was, succeeded in uniting the Sicilian Greeks in the face of the Carthaginian peril, a union that he made more lasting than any in Greece. And while Isocrates, the intellectual pure and simple, was idealizing another prince, Evagoras of Cyprus, Xenophon, with his experience of exile, travel, and war, was painting in his *Hieron* the portrait of the perfect monarch, whom he thought he had found in flesh and blood in the person of the king of Sparta, Agesilaus. He completed the exposition of his views by showing in the *Cyropaedia,* by the example of Cyrus the Elder, what the education of a prince should be and the benefits that his people can expect as a result. Isocrates finally came back to reality, a reality that endangered his country's liberty, when he decided that he had found in Philip of Macedon the man of destiny who could bring Panhellenism out of utopia into the realm of reality.

The desire for a philosopher-king

Philip had set his own statue beside those of the gods and thought of having the Greeks worship it in the sanctuary of Olympia. Alexander, persuaded of his divine parentage, instituted the cult of the ruler and demanded that he be made a god while he was still alive.

The arts and architecture. The same effort to escape from tradition, the same tendency to break the bounds of the city in order that the individual might forget his sufferings, showed in the arts in 4th-century Greece. But the main principles held firm, and the combination of old and new still produced marvels of art, closer to man and his sufferings than in the past. Artists conveyed the uncertainty of a world in transition. Men sought to escape the anguish of the fleeting present by creating an ideal world. Their attention was concentrated less on the solid past, sometimes put in doubt today, than on daily life, even among the humble and the have-nots, and on the human personality and the passions that shake it. Thus developed the taste for analysis.

Developments in sculpture

The sculptors grew away from the city and forgot the existence of the gods. They felt the influence of painting, which renders shadows, reflections, and transparencies and translates the movements of the soul, and they set themselves to study man, no longer in what is universal in him but in his fears and weaknesses. Hence, they sought to reproduce the expression of the face and of the eyes down to the finest shades of meaning, as well as interpreting gestures and bodily attitudes of suffering or emotion. Scopas expressed the pathos of existence in heads looking backward, in mouths twisted or half open, in quivering muscles. Praxiteles, more sentimental, interpreted voluptuous revery through the sweetness of a smile of invitation or resignation; his subjects no longer have the strength to hold themselves up, their bodies drooping softly and supporting themselves on a column. He was interested in languid youths and dared to bare completely the female body, even of goddesses, who appeal to the pleasures of the senses.

A similar grace, deliberately indiscrete, inspired toreutic art and still more the delightful figurines of Tanagra, little girls or young women captured in the very moment of a light thought or passing fancy, now flirtatious and laughing, now sad and melancholic, wrapped in transparent draperies or sauntering beneath a parasol.

A new movement developed in the second half of the century. Lysippus turned from women to athletes and was also interested in monarchs. Philip lured him to Macedonia, and he accompanied Alexander to Asia; he was the first court artist. But there is nothing formal about his art. He has an eye for detail, seizes the act of movement, and shows man as he is, in his trade or in history.

In architecture, novelty abounded. Theatres multiplied, and stone tiers made their appearance at Syracuse and Epidaurus, arranged in conformity with laws whose secret is lost today, to produce unequalled acoustics. In the religious sanctuaries, round buildings developed, *tholoi,* set above a crevasse giving access to the divinities of the underworld. Religious feeling remained strong in the Greek world wherever it was the expression of a community spirit. Cities and individuals gave considerable sums to reconstruct the great Doric temple of Apollo at Delphi, destroyed in an earthquake in 373. Later, thanks to Philip and Alexander, who were perhaps more interested in propaganda than in expressing genuine religious faith, Olympia was enriched with new buildings.

Ionian architecture

It was in Ionia, however, that architecture particularly flourished; it aimed both at ornament and at grandeur. Everywhere temples were built or rebuilt—at Sardis, Priene, Miletus, Xanthus. Indeed, on the Asiatic shore of the Aegean, already the home of the Ionic order, two of the Seven Wonders of the World were created: the temple of the goddess Artemis at Ephesus and the giant Mausoleum at Halicarnassus, the resting place of Mausolus of Caria, to whose apotheosis both Scopas and Praxiteles contributed.

Literature and oratory. In literature even more than in art, Athens remained the great capital of the Greek world. Writers who were not from Athens came there to learn its lessons. The Athenian political regime offered an example

of what not to imitate, but in literature, as in art, the age of masterpieces was not yet over. But the masterpieces were no longer inspired by the same spirit as in the previous century, which had seen tragedy, comedy, and history blaze into glory.

Except in the theatre, poetry, in decline, gave way to prose. As was natural at a time when man was more worried about his future than about his destiny, the output of drama continued, especially where the theatre was under the protection of tyrants, but the tragedies of this period have not been preserved. It is known, however, that they were in the same line as those of Euripides, poet of anxiety and of passion. Aristophanes continued to write but produced a new style of comedy, either plunging into utopia to escape the harshness of the times or taking the direction that was to be that of New Comedy by depicting in realistic fashion the manners of everyday life, drawn particularly from the common people, and by subtly analyzing situations rather than individual characters.

In the lawcourts and the world of politics, the spoken word extended its domain, since it was necessary to work upon the citizens in their double capacity as judges and as masters of their country's destiny. The orators stayed in the city to defend their clients in court and influence the decisions of the people but left it to appeal to Greek opinion and put forward great ideas.

Isaeus remained the specialist in inheritance taxes. Lysias, as previously noted, was not simply a speech writer for the benefit of those who could not speak in public: a panegyric at Olympia gave him the chance of addressing all the Greeks. Isocrates went still further. Having made a fortune as a lawyer, he became a professor of the highest form of eloquence, that for state occasions, insisted on a broad culture for his disciples; and he left his school to urge Panhellenism, first upon the Greeks and then upon Philip of Macedon, when Greece became for him more important than Athens.

Demosthenes

On the opposite side, with a higher flame of eloquence, Demosthenes continued his incessant struggle against the rulers of Macedonia. By his public speeches and political addresses, he fought obstinately in defense of Athens, of its liberty and its brilliant past, with a settled ardour and a passion of logic that finally defeated the skilled talent of Aeschines. But he was to pay with his life for his courage in opposing the power of Macedonia.

Demosthenes' life and work are one with the history of his times. There remained, however, some professional historians. None had the force or greatness of Thucydides, although one is in danger of misjudging, on the basis of fragments or mere echoes, historians who are known mostly by name only, such as Theopompus, Ephorus, or that Ctesias who was also a doctor at the court of Susa. On the other hand, the immense output of Xenophon has survived complete. An Athenian who often, though not always, admired Sparta, he left an account of 50 years of Greek history, finishing the narrative of the Peloponnesian War that Thucydides left incomplete and continuing his researches down to the Battle of Mantineia, which reduced the cities of Greece to greater chaos than ever. Full of curiosity, more percipient than he is usually given credit for, though suffering by comparison with Thucydides as a historian and with Plato as a philosopher, he stirs and interests his readers by the diversity of his talents. His troubled life took him to the crossroads of the civilizations of Asia and Greece. He appeared on the battlefields of both Asia and Greece, then gave up a soldier's life to become a country dweller, horseman, and hunter. Faithful to the memory of his master Socrates, he created both biography with his life of Agesilaus and the historical novel with his *Cyropaedia,* which foretold the coming of Alexander.

Philosophy and political thought. The philosophers in their schools also, as was natural, invited men to reflection, whether it led subsequently to action or to meditation. All the major philosophical schools were influenced by Socrates, who had directed his powerful mind to search for the good and the beautiful—*i.e.,* the true—and who had drunk hemlock with calm courage to show the justness of his principles.

In Athens, Antisthenes taught that neither happiness nor unhappiness existed in themselves but rather are subjective states. The Cynics believed the happiest creatures are the savage and the animal, since they suffer from none of the prejudices burdening man. They are lucky enough to know nothing of honours, country, religion. In short, to attain true happiness, one should "lead the life of a dog." Such was the blessed state that Diogenes was to attain in his barrel.

If the Cynics are the forerunners of the Stoics, the Epicureans derive from Aristippus of Cyrene, theoretician of hedonism. For this teacher and his disciples, everything in human life goes back to the senses. Happiness lies in pleasure; thus, the greatest happiness is the greatest pleasure. The wise man, however, is not insensible to the quality of the pleasure, and it is in his interest to remain master of his passions. Regarding the love he felt for a courtesan, Aristippus summed up in a striking phrase the teaching of the Cyrenaics: "I take; I am not taken."

Plato and the Academy

Like Xenophon, Antisthenes, and Aristippus, Plato was first molded by Socrates; their master perhaps taught them in the last analysis to be men more than citizens. That does not in any way mean that Socrates' teaching was harmful from the point of view of the state. On the contrary, Plato, before turning to philosophy, had wished to serve the city, and all his life he remained haunted by the problem of politics.

Socrates' death turned Plato away from active politics. After this, he combined the experience of travel with reflection in an attempt to discover the ideal regime of a philosopher on the throne. As has been noted, he thought he had found it in Sicily. Disappointed in this, he expounded the theory of it, first in *The Republic* and later in the *Laws.* Realizing that individualism was the death of the city, he sought to locate sovereignty in an all-powerful republic, authoritarian enough to abolish classes, and later in pure reason, creator of unchangeable laws. Plato would have liked to give concrete form to his abstract theory and suffered all his life from his inability to put his ideas into practice. Nor did he have the pleasure of seeing certain of his disciples write constitutions for cities after his death.

Since the regime and policies of Athens denied him a career as a statesman, he was a philosopher, having as his aim the formation of men for political life. On his first return from Sicily, in 387, he founded the Academy, a school of political science, where the listeners worked together under the direction of the master, aiming toward an ethical and scientific training whose ultimate aim was the reform of the state. Without ever repudiating Socrates, Plato went beyond him, broadened the field of reflection, and speculated on the realm of the absolute. Surrounded by his disciples, the most brilliant young Athenians of their generation, who live in his dialogues alongside Socrates, and in the company, too, of the Sophists, whom he delighted to confound, Plato sought passionately after truth in all fields, from mathematics to metaphysics. His dialectical method came to him from Socrates. It allowed him, through questions often tinged with irony, to lead his partner in the conversation by stages to the discovery of the solution. Often the dialogue resolves itself into connected exposition, and, when only imagery can transmit the subtlety of the supreme reality, Plato, the antagonist of poetry as an agent of decadence in the city, himself has recourse to the poetry of myths.

Socrates convinced Plato of the reality of virtue, and Archytas of Tarentum and the Pythagoreans showed him that the universe was founded upon the ever-new beauty of numbers. He knew that the world was a mere illusion. Thus, it is desirable to free oneself from one's mortal body and—thanks to the soul that survives it and that, by virtue of the reincarnations, remembers its past lives—to climb from illusion to belief and from belief to the world of ideas, the only reality, since it is the image of God. Borne upon the wings of love, the soul finally achieves its happiness in the divine contemplation of the beautiful, which gives it no longer belief but actual knowledge of the good and the true.

Plato goes further and much higher than Socrates, and it has been said that all subsequent philosophies are merely footnotes of his thought. He retains, however, the secret he

learned from Socrates of always providing the fascinating spectacle of a thought in the process of being born. At the same time realist and mystic, he is the father of the New Academy and of Neoplatonism. He opens the way for the Fathers of the Church and for Christian thought. The first in time of his famous disciples is, however, Aristotle.

Aristotle and the Lyceum

Aristotle had less charm than Plato but is more accessible. Where Plato opened up the lofty road of the idea, Aristotle opened up the earthly track of the fact. The two, however, make one whole—unique, incomparable, and profoundly Greek—in the sense that they settle forever the two complementary aspects of man's speculation.

Aristotle was born at Stagira, in the region of Chalcidice coveted by Philip of Macedon. His life, like that of Demosthenes, spans the middle of the century, from 384 to 322. For 20 years he was Plato's disciple, until the latter's death in 348/347. Five years later he accepted Philip's offer to take charge of the education of the young Alexander, then aged 13.

Aristotle, after living in Macedonia and touring Asia, founded a school known as the Lyceum at Athens in 335. There, scholarly discussion took place while walking, whence the name of "Peripatetic" for the school. In the afternoon the lectures were public. The morning was for work, in which Aristotle directed teams of his disciples. This is the only part of Aristotle's work that survived. For only the notes, fortunately copious, taken by Aristotle's pupils give us an idea, admittedly extensive, of the body of his work.

He left to Plato the realm of mathematical speculation but took in the rest of knowledge and established an encyclopaedic body of research on a combination of observation and syllogistic logic. In his eyes concrete objects alone exist, though they are constantly changing. He classified them, both flora and fauna, by genera and species, assembling an astounding mass of observations and conclusions. For him the life of species is determined by the action of a final cause, and motion is transmitted by a chain of events whose origin goes back to a god, perfect, unique, and indifferent to the world, though not totally divorced from the world, since he passes on his energy to matter whose forms he shapes.

In ethics, Aristotle is distinguished from Plato by his realism, as the latter distinguishes himself from Socrates by his idealism. Aristotle considers that virtue lies in a just mean, a sort of equilibrium that is obtained, like the pleasure of watching a tragedy in the theatre, by a purging of the passions. But, before being a moral creature, man is a political animal, and the first of all skills is political skill. Here again, though influenced by Plato, Aristotle differs from his master in not constructing an ideal city. He studies actual existing constitutions and arranges and classifies them as he classified animals and plants, following the same method that allows him, through critical observations, to work out definite laws for the great literary genres.

It was natural that a mind of this quality, this universality, should have left its stamp on Alexander; it was natural, too, that so exceptional a pupil should have shown his gratitude by sending his master money for his research and rare specimens for his collections from the depths of Asia. The Greeks soon saw that the young disciple had replaced encyclopaedic research with conquests up to the limits of the unknown, and substituted for the spirit of inquiry a determination to impose balance and order on conquered humanity.

The Hellenistic Age

THE CONQUESTS OF ALEXANDER THE GREAT

Relations between the Greeks and Macedonians. When, at the beginning of the summer of 336, Philip II of Macedon fell beneath the assassin's knife, there was an explosion of joy in Greece: the national enemy was dead. As for his successor, Alexander, the Athenians could believe that he was too young—a mere 20 years old—to follow in his father's footsteps. They knew that he was brought up on Greek and trained by Aristotle. They could also know that, before taking part in the Battle of Chaeronea at Philip's side, he had studied passionately Homer, Pindar, Herodotus, and Euripides; they could suppose that his encyclopaedic knowledge might possibly direct him toward the things of the spirit.

Alexander's background

Alexander had inherited from his father a taste for warfare, however, and from Homer and Herodotus he had drawn warlike lessons, as well as the example of a great struggle to be led against the barbarians of Asia. He had also listened to the advice of the orators; he knew of the crusades preached by Gorgias, Lysias, and Isocrates. He had not forgotten the march of the Ten Thousand across Asia nor the campaigns of Agesilaus; he knew that the Greeks had for too long allowed their quarrels to be settled by the barbarian king. As a Hellenized monarch, Alexander could believe he had every chance of bringing the Greeks finally into agreement beneath the dominion of someone not a barbarian. His father had shown him the way: Philip had created the League of Corinth in order to impose a general harmony under the aegis of Macedonia. Except for Sparta, all the Greek cities, members of the new league, had accepted in the spring of 337 the idea of a great war of revenge against Persia, by means of which Philip meant to cement their union.

Alexander took up Philip's program but turned it around: waging war in Asia was for Philip a means of ensuring his own sovereign authority over the Greeks; for Alexander, animated by an appetite for glory and for conquest, the war was itself the end. But, to attain such an end, he must necessarily have peace in his rear. His first act, despite his youth, was to intimidate the Greeks: he descended like a thunderbolt upon Thermopylae, haunted by the spirit of Leonidas, and won the succession to his father's seat on the Amphictyonic Council, whereupon the League of Corinth named him commander in chief of the Greeks.

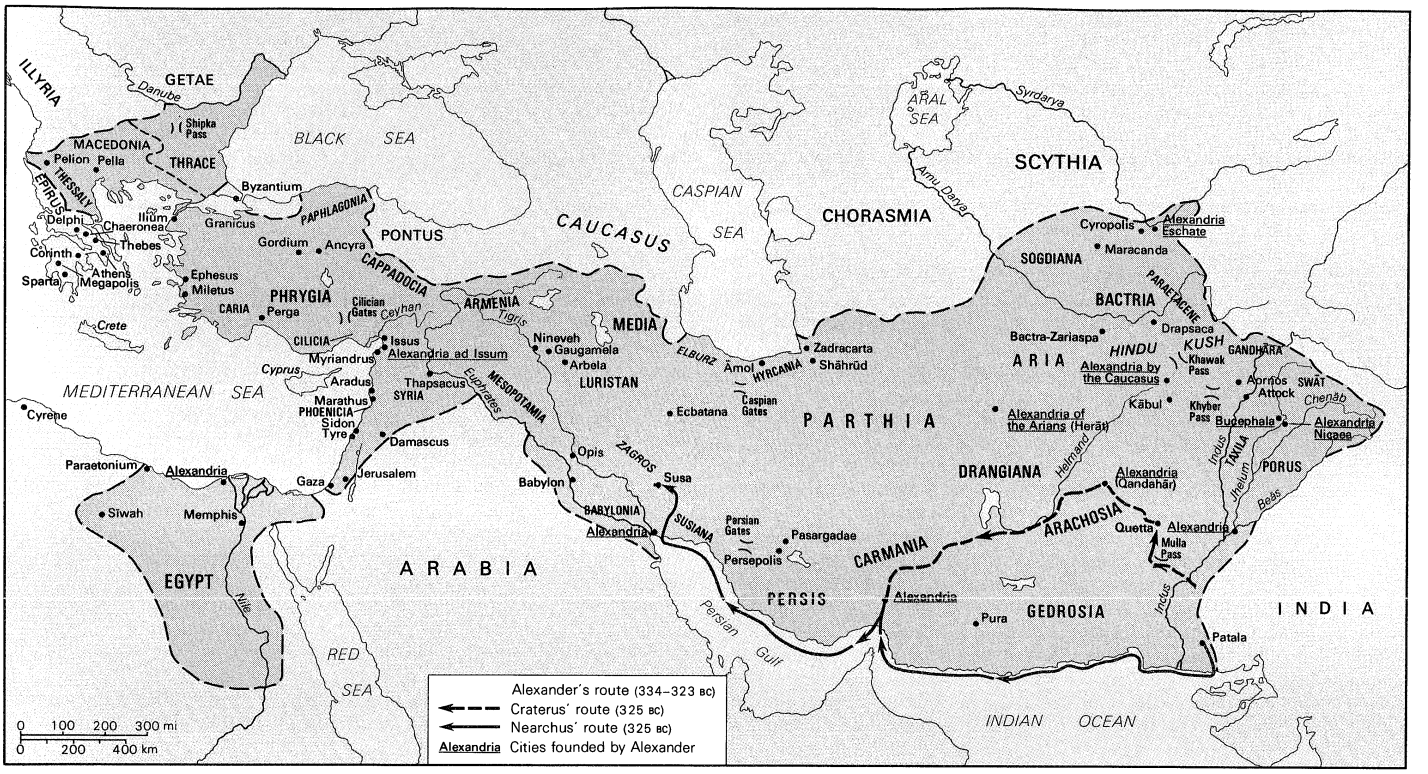
A danger now became apparent on the other flank, in the north, in 335. There was trouble in Thrace, and beyond, among the Triballi and the Illyrians. Alexander, with his army, reached the northern bank of the Danube and reestablished Macedonian order among the recalcitrant barbarians. Nevertheless, when a rumour spread that he had died in Illyria, the effect was seen of the embassies and of the gold that the King of Persia had sent into Greece to check the policy of Alexander's father and of Alexander himself. At Thebes the exiles returned and dared to kill the Macedonian governors. Thebes requested the support of the Arcadians, the Eleans, and, in particular, the Athenians, among whom Demosthenes was keeping alive the sacred flame of the struggle for liberty. They decided to send help secretly into Boeotia. But Alexander lost no time in showing that he was truly still alive.

From Illyria, he made a dizzy swoop into Greece: 12 days sufficed for him to appear before the walls of Thebes. The Thebans were dumbfounded, and the city fell. An example had to be made of them. The city was razed; its inhabitants, including the women and children, were slaughtered or sold into slavery. The punishment was so cruel that Alexander was struck with repentance; he made the pilgrimage to Delphi to beg Apollo's pardon and to cleanse himself in the god's presence from the stain of this bloody repression. He had commanded that naught be spared except the descendants of Pindar and the poet's own house. The terror was effective: the Greeks saw that it did not pay to take such a master lightly, even if he was flexible enough to measure out his vengeance sparingly, showing himself less severe toward Athens, a city for which he had always had a weak spot.

Destruction of Thebes

Now undisturbed either by barbarians or by Greeks, Alexander could turn toward Asia. He planned to make a bridge between the barbarian East and the civilization of the West, a project that presupposed replacing the Achaemenid Empire by a Macedonian monarchy. In the spring of 334 he took charge of a small but experienced army of 30,000 infantry and 5,000 cavalry and set off for the Hellespont. He had left behind his turbulent mother Olympias, who hated Antipater. But he could count on his faithful lieutenant to ensure the regency of Macedonia and the government of Greece.

The conquest of the Persian Empire (334–330). One might speculate about Alexander's aims as he crossed over



Alexander's empire at its greatest extent.

Adapted from *Westermann Grosser Atlas zur Weltgeschichte*; Georg Westermann Verlag, Braunschweig

from Europe into Asia. Did he intend merely to repeat the Persian Wars? to liberate from the Persian yoke the Greek cities of Ionia? to sow in the East the benefits of Hellenism? His ideas were large enough to take in all these goals, but it is certain that his sense of reality told him that he must crush Persia in order to put an end to the interventions of its arrogant diplomacy and hidden subsidies in the affairs of Europe. The hour had come for Macedonia to play the game up to the limits of the world. Alexander realized clearly what was necessary from a political point of view, but he also felt beating in his breast the heart of one of Homer's heroes; he wished to bring the *Iliad* to life.

Once he had crossed the Hellespont, Alexander made a sacrifice on the tomb of Achilles, whom he believed to be his ancestor, and then impetuously attacked the Persians. In 334, on the banks of the Granicus, after a crazy charge without a battle plan, he won a brilliant victory over the lieutenants of the Persian king. From the spoils of the barbarians he sent trophies to the Athenians that they might be consecrated in the Parthenon and adorn the Acropolis. He liberated the Greeks in the region of Troy from the King's yoke and exempted them from all the taxes imposed by his satraps. In a few months Ionia, land of the bards and fatherland of Homer, was conquered. At Sardis he restored to its inhabitants their ancient laws. At Ephesus he pacified the rival factions and gave the funds necessary to rebuild the temple of Artemis, one of the Seven Wonders of the World. Returning to northern Phrygia, he reached its capital, Gordium, on the road to Susa and there, according to legend, cut the famous Gordian knot that promised the empire of Asia to whoever should loose it. Alexander then moved south to cut Persia off from the Mediterranean, but in all the territory from the Hellespont to the Nile he wished to appear more liberator than conqueror and used force only where generosity failed.

Battle of Granicus

After experiencing serious difficulties in Cilicia in the autumn of 333, he penetrated into Syria and this time defeated the King himself in November at the Battle of Issus. Darius succeeded in escaping and reached Babylon, but he abandoned his dependents, his harem, and his family. Alexander respected his captives, and, when the Queen died in his camp as a prisoner, he ordered a funeral so magnificent that Darius considered leaving his throne to the Macedonian if he were finally defeated.

There was, however, no question of accepting a ransom or of negotiating. Alexander went his own way. Instead of making for Babylon, however, he completed the division of the empire into two halves, north and south, in order to break its naval power. Sidon opened its gates, and Tyre fell after seven months. Alexander escaped an assassin's knife and seized Gaza. Without a blow struck, Egypt fell into his hand and welcomed him as liberator. In the winter of 332/331, he offered sacrifice at Memphis, founded Alexandria, which was to be the cradle of Hellenistic culture, and advanced across North Africa toward Carthage.

Alexander next turned back toward the very heart of Persia. At Tyre he offered sacrifice to Heracles and also organized in passing a dramatic contest in the Athenian manner. He received from Athens the homage of an embassy that landed from a state trireme, charged with obtaining the freedom of the mercenaries from Athens whom he had captured at Issus serving in the Persian army. Since Athens had just refused its support to the king of Sparta, Agis III, who was trying to subvert the Greeks against Macedonia, ruled by Antipater in the absence of its king, he granted the Athenian prisoners their liberty. After this curiously literary and diplomatic interlude, he took up arms again, crossed new deserts, and passed over the Tigris; and on October 1, 331, at Gaugamela, not far from Arbela, he crushed Darius in a third and decisive battle. At the same moment, in the Peloponnese, in front of Megalopolis, Antipater was annihilating the forces of Agis.

Battle of Gaugamela

Reassured as far as the Greeks were concerned and now conqueror on the Asiatic front, Alexander made a magnificent entry into Babylon. A whole month of festivities allowed the soldiers to relax and enjoy themselves. Victors and vanquished celebrated alike, with a feeling of relief that the Achaemenid Empire had come to an end. The complete success gave substance to the idea of one single new empire beneath a monarch who abolished frontiers to try to bring about a fusion of East and West.

Instead of following Darius, who had taken refuge in Ecbatana, Alexander advanced toward Susa, which opened its gates before him. To symbolize the end of the age-long hostility of Greeks and barbarians, Alexander adopted

Persian costume and beliefs but at the cost of alienating his own men. He celebrated games, offered sacrifices, and freed the captive princesses, the mother and daughters of Darius, together with the young prince Ochus; he granted them Susa as their place of residence and had them taught Greek.

Advancing into Iran, he entered the holy cities of Persepolis and Pasargadae. He punished the resistance of the satrap Ariobarzanes by giving Persepolis over to pillage (apart from the royal palaces) and allowed his soldiers unprecedented loot. For his own part, he seized a fabulous treasure of 120,000 talents. At Pasargadae, he went to meditate before the tomb of Cyrus the Great and listened to the message of the dead conqueror. He had to adopt Eastern habits if he was to take the place of the Great King, even if it displeased part of his own army.

In the spring of 330 he organized at Persepolis an immense banquet, accompanied by music, songs, and dances. Abandon seized everyone. In a moment of exaltation, the son of Olympias, egged on by the courtesan Thais, mistress of Ptolemy, the future king of Egypt, set on fire the palace of the kings of Persia, which until that time had been spared. Darius, grown old, exhausted by his long flight, finished by falling into the hands of his satrap of Bactria, Bessus, who coveted the throne. He was put to death by the traitor; Alexander did not arrive in time to save him. He wept over the body of his enemy, had the last honours paid to him, and eight months later killed Bessus, whom he had taken prisoner. It was the end of the Achaemenids. There remained only one king of Persia, Alexander.

Death of Darius

The conquests in Bactria and the Indus Valley (330–323). Once Darius was dead, Alexander had the right to conclude that he had brought to a victorious end his mission as champion of the Panhellenic idea. He ceased to consider himself the delegate of the League of Corinth; he was no longer commander in chief of the Greeks but King of Kings, the successor of Cyrus the Great. He adopted the ceremonial of the Persian court, took a harem, and wore the tiara—a posture that led some to surmise that one Oriental despot had been exchanged for another.

Alexander went so far as to proclaim himself a god and demand from the Greeks abasement in his presence. In the army, anxiety gave place to anger. Plots were formed, and those suspected of complicity were executed, even if they were the King's old friends.

At a victory celebration in January 327 Alexander noticed Roxana, the daughter of his vanquished enemy; he eventually married her, thus realizing more than symbolic union of Greece and Asia.

There is no stopping on the high road of conquest. Mysterious India attracted Alexander because it held the traditions of his ancestral gods Dionysus and Heracles and because a pupil of Aristotle could there satisfy his curiosity as a geographer and his urge for exploration. What is more, the bearer of civilization would open up for the Greeks new commercial routes to the East, and his soldiers would restore their morale in guerilla warfare and high life alternately. The great adventure went on.

With an army of 120,000 infantry and 15,000 cavalry, he marched for 18 months until he reached the Beās (Hyphasis), an eastern tributary of the Indus. He suited his movements to the difficult country and the unknown climate; he adapted his tactics to the elephants of Porus, rajah of Lahore; he built a small fleet to sail on the Indus. In July 326 he finished off the resistance of Porus on the banks of another tributary of the Indus, the Jhelum (Hydaspes).

The final stage was to be the Ganges, at the very limits of the world, which scholars and legend said was bordered by the ocean. But his soldiers were upset by the tropical rain, his officers suffered from homesickness, and the army would go no farther. It was a bitter disappointment for the conqueror; he could not go on alone, and to the joy of the others he gave the order to return.

At the end of the autumn of 326 he launched a fleet, this time of 2,000 ships. Himself convalescing from a wound, he descended the Indus—a river that, instead of flowing into the Nile as was believed, has its own delta, as was discovered in July 325, and flows into the ocean. To reach Susa, the army was divided into three sections. In the north, Craterus was to march by the land route across the country of Arachosia; in the south, by the sea route, Nearchus would sail with the fleet; Alexander chose for himself the central route, across the fearsome desert of Gedrosia, in the hope of maintaining contact with the fleet.

The northern army made its march without incident. The parallel march by the King across the desert and the voyage of Nearchus, braving a terrible sea with a river fleet along an unknown shore bristling with reefs, nearly ended in disaster. Human energy won the day. Army and fleet first made contact again, against all hope, in Carmania; their meeting was celebrated in the autumn of 325 by an astonishing orgy lasting seven days. They met for good six months later at the head of the Persian Gulf in the delta of the Euphrates. They had thus opened up the sea route linking East to West, which the great sailors of the future would follow.

The survivors of the expedition were at last reunited at Susa. Alexander decided to reward them and at the same time to crown his policy of joining the races. In February 324 he organized an astounding marriage celebration at which the sound of the trumpet orchestrated all the movements. Ten thousand mixed marriages took place at the same time between Greek soldiers and Asiatic women; all the King's lieutenants received chosen Eastern brides. He himself, who had lost a son of Roxana in India, celebrated a new marriage with the eldest daughter of Darius, the princess Barsine (Stateira).

Greco-Persian marriage at Susa

After this he could disband part of his army, those who were tired of the country; he replaced it by native troops to whom he taught the Greek language and Macedonian tactics. He then incorporated them into what remained of his old troops. He did not, however, forget Greece and wished to bring about a higher union: having, as he thought, obtained a reconciliation in Asia, he had to impose it on Greece. He signed two decrees at Susa and published them at the Olympic Games in September 324. By the first, he ordered the Greek cities to recall their exiles, and all accepted this general amnesty except Athens, where the anti-Macedonian spirit was still alive. By the second, he demanded to be accorded divine honours; the Greeks recognized him as the new Dionysus, brother of Apollo and son of Zeus. In 323, at Babylon, his sovereignty became that of a king and god; the Greek cities of the entire world sent their representatives to celebrate his apotheosis.

Alexander's ambition was insatiable. He envisaged new explorations and studied fresh conquests, from the mouths of the true Nile to the Pillars of Hercules, Africa following upon Asia. But Nemesis lay waiting for him. His body had suffered too much from orgies, from wounds, and from the trials of war. Fever took him, increasing daily, and on June 13, 323, death struck the god who was not immortal.

The empire and achievements of Alexander. It is difficult to say what would have become of the empire of Alexander if it had been given him to reach the Ganges and finish his work. One's judgment must limit itself to what this exceptional man tried to do and to the trends that had already manifested themselves.

He did not live to see the known world united under his sovereignty, nor did he give the world a single capital. Greece stayed Greek, with its internal conflicts, beneath the rule of Antipater, Alexander's representative in Europe and governor of Macedonia. Asia remained Asian without having imposed upon it great innovations. The conqueror had the intelligence to preserve all that was workable and healthy. He was content to correct or refine the Achaemenid administration. He reorganized the satrapies but kept them in existence in Syria, in India, in Egypt: he gave each satrap a deputy, with a general as military commander. He set in order the payment of taxes, since the treasures of the kings of Persia were not inexhaustible, and issued a new coinage that covered the known world with his portrait. There was no area that did not bear the master's mark: justice functioned; circulation of money was encouraged; canals were dug and roads opened or repaired; town life and agriculture throve alike.

The master, king and god, might have chosen Babylon

as capital, but he remained the pupil of Aristotle. He believed that it was the task of the West to enlighten the East, to stir up its masses, to energize its softness and delays. He protected artists, scholars, poets, and writers trained at Athens, whence came his admiration for the capital of the arts and of literature. Knowing by experience the vital strength of Greek language and civilization, he founded across the world 34 Alexandrias, future centres of Hellenism, on African or Asian soil. It is from Alexandria in Egypt and the Alexandrians that the world possesses all that Greek civilization from Homer onward had bequeathed.

Such indeed was the master plan of this young genius, and it succeeded after his death. As if he had realized that the immediate question of the succession would cause internal strife, he sought to ensure for the generations to come the survival of Hellenism. He went far beyond the utopian and superficial Hellenism preached by Isocrates, which Philip had almost brought about. If Alexander imposed his monarchy by divine right upon the world that he had made to yield or had conquered, it was not to leave side by side brothers who were still enemies or races separated by mutual hatred; it was in order to mold these divers elements into one single unit and to pour the best of Greek thought into an Eastern mold, where authority from on high would in no way derogate from the freedom of equals or inferiors.

When he married Roxana, when he gave his lieutenants Eastern princesses for wives, when he gave his soldiers Persian women, he was only setting an example or rather mildly demanding one. He meant to bring about a marriage between two civilizations, and he was certain that the fruit would have the best elements of both. Dead at the age of 33, after 12 years of action, Alexander had made history of a new sort, that of the conqueror who brings peace, who pardons the vanquished, who forces nations to accept his rule, only to help them to live and to unite them.

Ideas rule the world. Alexander passed on his ideas to the world that he had conquered. Before him, neither Sparta nor Thebes had been able to dominate Greece by arms; even Athens had not succeeded, either by force or by intelligence. With Alexander the rule of factions was over, the cities were broken, leagues were dissolved, nations cast into the melting pot, and races intermingled, and the time of great men itself was over.

While Aristotle was drawing up a catalog of the products of earth and man and was building up a systematization of the Greek world, his Macedonian pupil was saving the Greek heritage, was opening the Alexandrian Age, and was allowing Hellenism to diffuse throughout the world.

(E.D.)

THE HELLENISTIC AGE (323–30 BC): GREECE IN THE ROMAN PERIOD

The establishment of the Hellenistic kingdoms (323–276 BC). Alexander's career gave rise to a new Greek world, in which the city-states played a role subordinate to territorial monarchies but in which the genius of the Greek people found fresh outlets in every field of culture and political experimentation. The three centuries from Alexander to the establishment of the Roman Empire under Augustus were among the most productive and influential in the whole of Greek history.

From Alexander's death to the death of Perdiccas, 323–321. Alexander's death (323) produced a grave political crisis over the succession; in a compromise solution the army vested power jointly in Philip's feeble-minded bastard, Arrhidaeus, who was an epileptic, and the unborn child of Alexander and Roxana, should it prove a boy (as it did). These two became Philip III and Alexander IV. Powers and provinces were apportioned. Antipater was confirmed in his control of Macedonia and the European possessions, including Greece. In Asia Perdiccas was made chiliarch (or grand vizier); Craterus was declared guardian of the kings. Ptolemy received Egypt (where Cleomenes of Naukratis had usurped power) and gained prestige by seizing Alexander's body, which was being conveyed to Macedonia, and diverting it to Alexandria. Antigonus Monophthalmus (the One-Eyed), satrap of Phrygia, received western Anatolia (Greater Phrygia, Lycia, and Pamphylia), and Lysimachus received Thrace (hitherto an appendage of Macedonia). Alexander's Greek secretary, Eumenes of Cardia, was sent to oust Ariarathes from Cappadocia and Paphlagonia. The satrap of Hellespontine Phrygia, Leonnatus, soon perished bringing reinforcements to Greece during the Lamian War. Of all these men Ptolemy, Antigonus, Eumenes, and Lysimachus were the most outstanding and played the greatest part in the struggles for power that quickly ensued.

Political crisis over succession

News of Alexander's death provoked a rising in Greece known as the Lamian War (from the siege of Lamia). The two main disaffected factions were, first, the Athenian democrats under Hyperides, who voiced the widespread resentment felt at Alexander's decree recalling exiles, and, second, a body of mercenaries back from Asia who had provoked disturbances at the mercenary market of Taenarum in Laconia and, under the Athenian Leosthenes, now brought help to Athens. Aetolia and Thessaly joined the movement, soon followed by other Greek cities. Their forces besieged Antipater in Lamia; but Leosthenes fell in battle, reinforcements reached Antipater, and the Athenian fleet was crippled in a defeat at Amorgos. In 322, under pressure from both Antipater and Craterus, Athens negotiated; in the peace it lost Oropus and had to pay an indemnity and accept a Macedonian garrison in Piraeus. Both Hyperides and Demosthenes perished in the reprisals, and an oligarchic government was established under Phocion and Demades. Greece was now bound more tightly than ever to Macedonia, without even the pretense of alliance that the League of Corinth had offered. Aetolia alone escaped because Antipater and Craterus were recalled to Asia.

In Asia, in Craterus' absence, Perdiccas had usurped guardianship of the kings and was making a bid to unite the empire under his control. Antigonus of Phrygia hated Perdiccas; and personal issues exacerbated the conflict. Though married to Antipater's daughter Nicaea, Perdiccas was tempted by the offer of the queen mother Olympias to give him her daughter Cleopatra, Alexander's sister. Perdiccas' equivocations helped provoke a coalition of hostile satraps: Antipater, Craterus, Antigonus, Lysimachus, and Ptolemy all united against him. But war was averted when Perdiccas, invading Egypt, was assassinated (321). Ptolemy, who had had the usurper Cleomenes murdered, was now firmly in control there and strengthened by the addition of Cyrenaica to the west; he was wise enough to decline an invitation to take Perdiccas' place. Perdiccas' death brought the first period of manoeuvre to an end. It also isolated Eumenes, who was closely linked to Perdiccas; but a drive against him proved a fiasco, though it brought about Craterus' death.

Coalition against Perdiccas

The career of Antigonus Monophthalmus, 321–301. The two decades following Perdiccas' death were dominated by Antigonus' attempt to succeed him as the unifier of Alexander's empire.

A meeting of Perdiccas' opponents at Triparadisus in north Syria (321) declared Antipater guardian of the kings in his stead. Babylonia was assigned to Seleucus, one of Perdiccas' murderers. Antigonus was entrusted with the task of hounding down Eumenes, whom the allies had condemned to death; having appointed Antigonus general of Asia, Antipater returned to Europe with the kings.

During the next two years (321–319), in the drive against Eumenes, Antigonus rode roughshod over his colleagues in Asia Minor and was virtually master of the whole area. Eumenes was penned up in Nora in Cappadocia, when Antigonus learned of Antipater's death (319). Antipater had arbitrarily appointed as regent to succeed him Polyperchon, an old officer of Philip II, instead of his own son Cassander. Cassander, an able and ambitious man, promptly crossed into Asia, there to organize a coalition of Lysimachus, Antigonus, and Ptolemy against Polyperchon. Seduced by these new hopes Antigonus rashly offered to give Eumenes his satrapy back in return for his collaboration—an offer Eumenes readily accepted without any intention of keeping his word. Meanwhile, Ptolemy, on hearing of Antipater's death, had launched an attack on

Syria–Phoenicia, an act less important for what it achieved than it was as an indication that he had now made both the kingdom and the strategy of the pharaohs his own.

In Europe Polyperchon, opposed by the other governors, attempted to improve Macedonian relations with the Greeks. Athens was promised the return of Oropus and Samos, and the events of the Lamian War were expunged in a formal proclamation; the oligarchies set up by Antipater were to be abolished and exiles restored. These gestures were halfhearted and short-lived. At Athens a brief spell of democracy (during which Phocion and others were executed) ended in negotiations with Cassander's troops in Piraeus as soon as it became clear that Polyperchon was losing ground.

An agreement made in 317 established the Peripatetic philosopher Demetrius of Phaleron as master of Athens, where he remained in power for the next ten years. Under his guidance Athens enjoyed peaceful and efficient government, in which some of the precepts of Aristotelian political thought were put into practice; the presence of Cassander's garrison stifled opposition. Polyperchon himself, meanwhile, had been expelled from Macedonia; having lost his fleet in a defeat by Antigonus near the Bosporus and with Cassander gaining ground in Greece, Polyperchon fell back on the Peloponnese, where he still commanded some support.

Power struggle in Macedonia

Meanwhile, events in Macedonia had ended all pretense of loyalty to the family of Alexander the Great. The quarrel between Polyperchon and the other dynasts had been reflected in the fortunes of the kings. Eurydice, the wife of Philip III, had joined Cassander and proclaimed him regent; and Polyperchon, who still held Alexander IV, now recalled Olympias from Epirus to strengthen his claims. Olympias had both Eurydice and Philip III killed; but she was herself imprisoned by Cassander shortly afterward, tried and condemned by the Macedonian army assembly, and executed. To improve his own position relative to Polyperchon, Cassander married Thessalonice, a half sister of Alexander the Great (and sometime later named the city he founded on the site of Therman after her).

Polyperchon had taken the obvious step of allying himself to Eumenes and as regent had assigned to him the generalship of Asia. From 318 onward Eumenes fought several impressive and successful campaigns against his united enemies—in Asia Minor, Phoenicia (where he recovered some of Ptolemy's conquests), Babylon (which he took), and Persia. But in 316 in Persia he was betrayed by his own troops, handed over to Antigonus, tried, and executed. Over several years Eumenes had employed skill and resilience in the name of the kings (though whether indeed with the selfless loyalty alleged by the ancient sources may be questioned). His death left Antigonus free to act against the satrapy of Iran, where anarchy was rife, and to pursue on his own behalf the unity of the empire. In so doing Antigonus became the rival and hated enemy of the other generals; thus Eumenes' death marked a crucial step in his struggle to win the succession of Alexander.

Antigonus gained control of the whole area from Asia Minor to Iran, and in 315 he turned southward to expel Seleucus from Babylonia. As a refugee in Ptolemy's court, Seleucus urged action on the other generals, and an ultimatum was delivered on Antigonus in north Syria by Ptolemy, Cassander, and Lysimachus. Because Antigonus' action against Eumenes had been on behalf of them all, they called upon him to restore Babylonia to Seleucus, to abandon Syria to Ptolemy, to hand over Hellespontine Phrygia to Lysimachus, and to surrender Cappadocia and Lycia to Cassander; further, they demanded that Eumenes' treasure be shared with the rest. Rejecting these demands, Antigonus continued his annexations; soon he held all southern Syria except Tyre, and he turned thence to seize Bithynia and Caria. Like Eumenes before him, he also allied himself with Polyperchon, whom he recognized as general of the Peloponnese.

Antigonus' manifesto of 315

In 315 at Tyre, Antigonus published a manifesto recording the condemnation of Cassander by his army assembly for the "murder" of Olympias and the seizure of Alexander IV and his mother; he also published his own appointment as guardian of the surviving king, thus stressing his role as defender of the legitimate line. Further, the proclamation declared that the Greek cities should be free, autonomous, and ungarrisoned—a demand to which Monophthalmus was to remain consistently faithful; he followed it up (probably in 315–314) with the creation of a confederation of the Aegean islanders (Nesiotes), engineered a revolt of Delos from Athens (it was to remain independent until 166), and sent agents to Greece to intrigue against Cassander.

Leaving his son Demetrius to guard Phoenicia against Ptolemy, Antigonus planned to attack first Lysimachus in Thrace, then Cassander in Macedonia. But Ptolemy, after some delay caused by revolts in Cyrene and Cyprus (where he had now gained a footing), yielded to the pressure of Seleucus and, in 312, attacked and defeated Demetrius at Gaza, thus enabling Seleucus to recover Babylonia and drawing Antigonus southward. Seleucus' change of fortune, however, suggested that negotiations might profit both Antigonus and Ptolemy (who had returned to Egypt), and after some preliminaries an uneasy peace was agreed (311).

This peace between Antigonus and Cassander and Lysimachus, on the one hand, and Ptolemy, on the other, represented a compromise based on the status quo. Cassander was to remain general of Europe during Alexander IV's minority; and Antigonus, general of all Asia. Lysimachus' control of Thrace and Ptolemy's of Egypt were confirmed. Seleucus, no party to the peace, was quickly excluded from all mention (he remained at war with Antigonus for two more years); nor did the peace concern Polyperchon. A clause underwriting the liberty of the Greek cities was destined to furnish each signatory with a convenient *casus belli* against any other who could be shown to have ignored it. In about 310 Cassander removed the limiting clause on his own powers by assassinating Alexander IV and his mother. For Antigonus the peace merely afforded a breathing space before his next attempt to absorb the other four territories.

The next ten years reveal a complicated story because the general alignment against Antigonus did not prevent his rivals from intriguing against each other or from making temporary accommodations with him. For both Antigonus and Ptolemy naval power was assuming increased importance; Antigonus had his eyes on Macedonia, and Ptolemy's control of Cyprus and Syria–Phoenicia turned on the possession of a navy. In 310 Ptolemy made his brother Menelaus general of Cyprus, and it was perhaps in 306 or even earlier that the Egyptian ruler entered into a long-lasting alliance with the maritime commercial city of Rhodes. Meanwhile, Seleucus recovered the upper satrapies, inflicted a defeat on Antigonus, and made a peace requiring Antigonus to surrender his claim to Iran but leaving him free to fight against the Mauryan king Candragupta in India.

About the same time, Cassander made a rapprochement with Polyperchon (who abandoned his support of Heracles, an alleged bastard of Alexander). That accord led Antigonus and Ptolemy to patch up an agreement, which may have been linked with Ptolemy's assassination of Polemaios, Antigonus' agent in Greece. The background is obscure, however, and Ptolemy's reason for this turnabout can only be surmised. The mainland Greek cities, faced with a coalition of Cassander and Polyperchon, appealed to Ptolemy, who invaded the Peloponnese in 308; but failing there, he came to terms with Cassander (though Ptolemaic garrisons remained installed in Corinth and other Greek cities). The following year (307), while Cassander was busy in Epirus, Antigonus' son Demetrius Poliorcetes sailed to Athens, seized it, and expelled Demetrius of Phaleron, who was superseded by a democracy. When Demetrius Poliorcetes seized Cyprus, however, the friendship with Ptolemy terminated. In 305 Demetrius began his famous siege of Rhodes, which lasted a year and ended in a compromise peace. Rhodes accepted the Antigonid alliance provided it entailed no disloyalty to Ptolemy; the outcome of the siege was celebrated by the Rhodian erection of the Colossus in the harbour at Rhodes and by the title of Poliorcetes (the Besieger), thenceforth borne by Demetrius.

Conquests of Demetrius Poliorcetes

The year 306 was marked by a spate of kingship claims.

After Demetrius' seizure of Cyprus, Antigonus assumed the title of king for both himself and his son—a clear implication that they were kings of Alexander's empire. A year later Ptolemy also took the title—no doubt implying "king of Egypt" (where he was already pharaoh to the native Egyptians). Cassander, Lysimachus, and Seleucus quickly followed suit. In proclaiming a Seleucid era from October 312 (April 311 in the Babylonian calendar), Seleucus had already perhaps taken a step in this direction; but his regnal years date only from 305–304.

Demetrius was now master of Athens; and after gaining the Isthmus region he resuscitated the Greek confederacy, with himself and Antigonus as presidents. Its constitution, known from an inscription found at Epidaurus, owed much to Adeimantus of Lampsacus. Like the league of Philip II, it was intended primarily as a tool of Macedonian domination. (Shortly afterward Polyperchon disappears from the records and was probably dead.)

Antigonus' enemies now prepared seriously to oppose him. A new coalition of Cassander, Lysimachus, Ptolemy, and Seleucus (who had ceded India to Candragupta in exchange for a troop of invaluable war elephants) made ready to confront Monophthalmus, who recalled Demetrius from Athens. Father and son met the combined forces of Lysimachus and Seleucus (with his elephants) at Ipsus in Phrygia (301). Antigonus lost the battle, perishing in the melee; Demetrius escaped.

The Battle of Ipsus and death of Antigonus

The victors shared the spoils. Lysimachus took Asia Minor to the Taurus (except parts of Lycia and Pisidia, which became Ptolemaic, and Cilicia, which was governed for a time by Cassander's brother Pleistarchus). Seleucus declared his claim to Syria, but he did not care to press it against his benefactor Ptolemy, who held all south of Aradus and Damascus. Cassander remained satisfied with his European possessions. The battle at Ipsus marked the end of all pretense that the empire would be united; despite Lysimachus' realm astride the straits, Europe and Asia were thenceforth to go different ways.

The failure of Demetrius Poliorcetes, 301–286. Though defeated at Ipsus, Demetrius still possessed the islands, his fleet, and several cities in Greece and Asia; but he lost Athens, which had been offended by his outrageous behaviour and had swung over to Cassander as a nominally free city under the government of Lachares. Demetrius' first chance came when Ptolemy made a double marriage alliance with Lysimachus; Seleucus, in consequence, turned to Demetrius (rather than Cassander) to make a counteralliance (at Rhosus in Syria), by which Seleucus married Demetrius' daughter Stratonice and Demetrius received Cilicia. It proved short-lived, however. Attempting to take advantage of Cassander's death in 297 (or possibly 298), Demetrius started in 296 to regain a foothold in Greece and retook Athens, which he held from 294 until 288; but he lost his remaining Asian possessions—Cyprus to Ptolemy, Cilicia to Seleucus, and his Ionian cities to Lysimachus. In 294, however, he successfully invaded Macedonia, killed one of Cassander's young sons, expelled another, and was proclaimed king of Macedonia by the army. Soon afterward he celebrated his new dignity by founding Demetrias in the Gulf of Pagasae (modern Gulf of Vólos, in Greece).

Demetrius' volatile temperament was inconsistent, however, with any firm policy. In 292–291 he invaded Asia, only to be forced to return to Europe by risings in Boeotia and Aetolia stimulated by Pyrrhus, the young king of Epirus whom Ptolemy had restored to his kingdom in 298–297 to maintain pressure on the Macedonian frontier. Demetrius suppressed the Boeotian rising and in 291–290 recovered Corcyra, which Pyrrhus had received as dowry from Agathocles of Syracuse, whose daughter Lanassa he married in 303. Demetrius' conflict with Pyrrhus went on for two more years, ending with a peace in 289. From then onward Demetrius' position deteriorated. By about 287 he had lost all the Aegean islands to Ptolemy; worse still, in 288 Lysimachus and Pyrrhus had combined to fall upon Macedonia and expel Demetrius, while Ptolemy's fleet liberated Athens. Thenceforth Demetrius was a wanderer without a base. He attacked first Lysimachus, then Seleucus, who took him prisoner in 285 and put him under house arrest. Demetrius, who had become an anachronism in a world of solidifying frontiers, died from drink in 283.

The expulsion of Demetrius

The end of the struggle for succession, 286–276. Antigonus' death at Ipsus had left Lysimachus the most serious challenger to the throne of Macedonia. He now held territories on both sides of the straits, where already in 309–308 he had built the stronghold of Lysimacheia. In 288–287 he had annexed northern Macedonia (having expelled Demetrius with Pyrrhus' help); and in 285, having broken with Pyrrhus, he seized the rest of Macedonia and Thessaly—though Demetrius' son, Antigonus Gonatas, still held Demetrias, Chalcis, and Corinth, later to be known as the "fetters of Greece" and certainly the bastions of Macedonian rule. To these for a time was added Piraeus.

Lysimachus was brought down at the height of his power, thanks partly to a domestic drama. In the interest of her children, his third wife, Arsinoe, Ptolemy's daughter, persuaded Lysimachus to execute his son Agathocles as a traitor. Thereupon Agathocles' widow, Lysandra, and her brother Ptolemy Ceraunus (Ptolemy's children by an earlier marriage) stirred up Seleucus to make a bid for Macedonia. Seleucus, perhaps also encouraged by Lysimachus' governor in Pergamum, who now went over to him, invaded Asia Minor in 282; Lysimachus was defeated at Corupedium and perished.

Seleucus, the last of Alexander's generation, seized Lysimachus' possessions in Asia, crossed into Europe, and was promptly assassinated by Ptolemy Ceraunus. Seleucus' death did not threaten his realm, where his son Antiochus I already reigned as co-monarch. The deaths of Lysimachus and Seleucus thus offered the greatest opportunity to Antigonus Gonatas, who could not, however, stop Ceraunus from being accepted as Macedonia's king. Ceraunus' reign, however, was violent and brief. To strengthen his claim to the throne, Ptolemy Ceraunus married Lysimachus' widow, Arsinoe (his own half sister); but when he murdered two of her sons, Arsinoe fled to Egypt and married her own full brother, Ptolemy II (ruled 285–246), who had succeeded their father. Soon afterward (probably in 280) Macedonia was invaded by a band of Gauls (Celts) migrating across Europe and quick to penetrate frontiers weakened by Lysimachus' defeat. Ceraunus perished at their hands, probably in January 279; a Gaulish kingdom of Tylis was established in Thrace, and one group reached Delphi, where the Aetolians brought about their destruction. Later other bands crossed into Anatolia.

Ceraunus' death was followed by short weak reigns in Macedonia and (apparently) by a period without kings at all. Gonatas, who had been weakened by his failure to check Ceraunus, faced revolts in Greece, where Sparta rose under Areus and where Argos and Megalopolis expelled his garrisons. But after a short campaign in Asia and the striking of what was to be a lasting alliance with Antiochus I (Gonatas married Phila, Antiochus' half sister), Gonatas returned to Europe to defeat the Gauls in Thrace at Lysimacheia. This success, widely publicized in the propaganda of the dynasty, enabled him to make himself master of Macedonia and Thessaly by 276; a king since 283, he had at last found a kingdom. Thus, in just less than 50 years from the death of its founder, Alexander's empire had settled down into the three territorial states that along with Pergamum were to constitute the great powers of the Hellenistic world until the Romans overthrew them: the Ptolemies in Egypt, the Seleucids in the composite state based on Syria and Mesopotamia, and the Antigonids in Macedonia. Those three states were the survivors of the dynastic struggles of the successors; and each dynasty had its own special problems.

The three great powers of the Hellenistic world

The Hellenistic monarchies in the 3rd century BC. *Ptolemaic Egypt.* Many Greeks had entered Egypt in the 4th century as mercenaries and traders; and, after Alexander founded Alexandria, their numbers were swollen by Greeks and by half-Hellenized Thracians and Anatolians who poured in as mercenaries.

Settled in the few Greek cities—Alexandria, Naukratis, and Ptolemais—or on plots throughout the countryside, where as cleruchs they constituted a reserve army and

there organized themselves in their own communities (*politeumata*), the Greeks built gymnasiums and wrestling schools, formed clubs, and read Greek books. This Greek minority formed the base of Ptolemaic rule. Alexandria was an exception: besides Greeks it contained Jews (only superficially Hellenized) and Egyptians; and, although it housed the palace and central administration, it was officially separate from Egypt and never a Greek city in the traditional sense. In the countryside, moreover, though encouraged by the early Ptolemies, Hellenism was a wasting asset; as the influx of immigrants declined, native pressures grew. Intermarriage produced mixed families, and the cultural traditions of uprooted Greek immigrants gradually gave way to the centuries-old social patterns of the native Egyptians. From about 200 BC onward Egyptian traditions began to prevail in all spheres. Nevertheless, so long as Ptolemaic rule endured, there remained a strong incentive to acquire the hallmark of Greek culture. Thus in Egypt as elsewhere, Hellene came to be a cultural term, valued especially because the Greeks remained the privileged class on whom the administration rested.

The character of Ptolemaic government was linked closely with foreign policy. The security of Egypt was assured by outposts in the eastern Mediterranean and Aegean and by an expensive naval supremacy. Southern Syria, seized in 301, remained Ptolemaic until 200. Cyprus, too, after some dispute with Antigonus I and Demetrius I, became permanently Egyptian. For most of the 3rd century, the Ptolemies controlled the Cyclades (and Samos to 201), much of the Asian coast from the Calycadnus to Ephesus, and much of the Hellespontine and Thracian coasts with Lesbos and Samothrace, together with outposts in the Peloponnese and Crete.

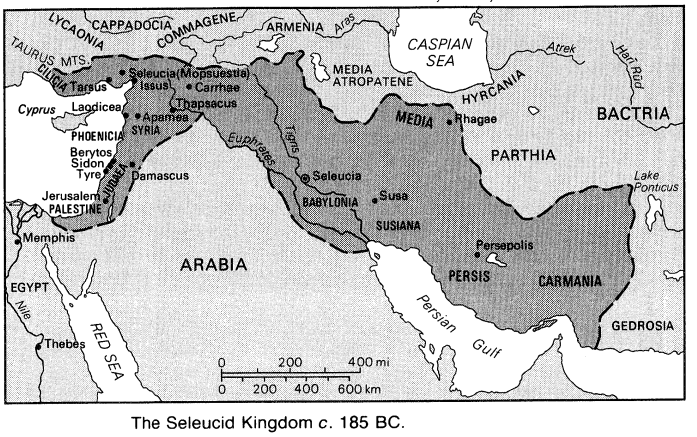
Though perhaps conceived defensively, the acquisition of the Ptolemaic Empire led to conflicts with both Antigonids and Seleucids (see below *Ptolemaic policy in the Aegean and Asia Minor*); nevertheless, it brought economic gain to the early Ptolemies, who pursued a mercantilist policy, perhaps primarily in support of their costly imperial program. Essential metals, timber, and manpower were lacking in Egypt, together with many of the immigrants' basic needs. Thus the Ptolemies were led to develop a highly centralized bureaucracy, which was the most individual feature of their kingdom.

From 200 onward the bureaucracy became less efficient, and a revival of Egyptian influence occurred in many fields—in the army after 217, in the adoption of Egyptian customs, and in the predominance of the Egyptian calendar. Nevertheless, Egypt was the last of the Hellenistic states to be absorbed by Rome (see below *Roman annexation of Egypt*).

Three distinct Seleucid regions

The Seleucid kingdom. The Seleucid dominions covered three distinct areas—Asia Minor (Anatolia), Mesopotamia (with Babylonia, Assyria, northern Syria, and, after Antiochus III's victory at Panium in 200, Coele Syria [Palestine]), and Iran. Their straggling and heterogeneous character created rival claims to attention, which made for weakness. Asia Minor contained a variety of states and peoples. The Seleucids controlled Hellespontine Phrygia, Phrygia proper, Lydia, inland Caria, eastern Cilicia, and for a time southern Cappadocia; several minor dynasts acknowledged them as overlords, but only the royal road to Sardis held the area together. Between the ancient trading and manufacturing districts of Mesopotamia and the Syrian and Phoenician coastal cities, which had long been exposed to Mediterranean influences, were the nomads, shepherds, or cultivators, including the Jews. In this predominantly Semitic area, with its worship of Bel, Marduk, Yahweh, and many Baalim, the Seleucids encouraged the revival of Babylonia, where the old religion and cuneiform writings underwent a renaissance. Farther eastward a mixed population of nomads, shepherds, farmers, and, in the cities of Media and Persia, traders shared an Iranian culture resistant to Hellenism. Failing there, the Seleucids gradually fell back upon the centre, where Antioch was to outlast Sardis and Babylon as the capital; but the Seleucid attempt to unite this disparate empire was an experiment not wholly fruitless.

By courtesy of John Bartholomew & Son Ltd.



The Seleucid Kingdom *c.* 185 BC.

The amorphous Seleucid realm was unified only by personal allegiance to the king and by the Hellenization he promoted. Alexander's foundations were supplemented both by full cities—often with dynastic names, like Antioch-on-the-Orontes, Seleucia-in-Pieria, or Laodicea-on-Sea—and by military settlements (which later sometimes achieved the status of cities), the inhabitants of which manned the phalanx. Many of the cities enjoyed nominal independence; the degree to which this was achieved in reality reflected the goodwill or weakness of the king. Antiochus III claimed full sovereignty over the Asian Greek cities; yet in 196 Miletus and Magnesia could engage in warfare. Many towns had a royal governor, and freedom often meant only freedom from tribute. (F.W.W.)

Syria became the centre of the Seleucid state and was the area most heavily influenced by the Greek culture of the Macedonians. In Babylonia, on the other hand, the traditions of the past lived on; the Mesopotamian cuneiform script, for example, continued to be used by the influential priestly caste, just as in pre-Seleucid times.

The ruling class in the Seleucid state was a Greek-speaking Macedonian aristocracy descended from the original conquerors. Below them was a class of Greek settlers, mostly soldiers, officers, and functionaries, located in the eastern provinces as well as in the Greek cities of Anatolia. Greek was spoken everywhere, and Greek schools were to be found throughout the realm. Inscriptions in Greek have been found even in the areas farthest removed from traditional Greek centres. In many places, Greek theatre, performed by travelling actors, became popular. Inscriptions of Euripidean verse have been found in Armenia, and Greek was used for legal purposes at Susa in Iran.

On the whole, the Hellenistic culture was limited to the cities; in the countryside the indigenous national characteristics as well as the native languages of the Near East survived.

Resistance to the Seleucids

Thus resistance to Macedonian rule was centred mostly in rural areas (*e.g.*, the movement that established the Parthian state in the 3rd century BC, and the rising of the Maccabeans in the 2nd century). Antiochus IV (175–164/163), in his zeal to propagate Greek culture, raised a statue of the god Zeus in the temple at Jerusalem. He was supported by Hellenized Jewish aristocracy, but beginning in 165, upholders of traditional Judaism rose against him under Judas Maccabeus. After 25 years of struggle the Maccabeans won Judaea from the rule of the Seleucids.

For the Macedonians and Greeks who left their homelands in great numbers during the 3rd century BC, a broad new field of activity opened up under the Seleucids. In their domains, Greek-speaking people banded themselves into organizations; in these clubs as well as in private life, the traditions of Greek culture and the Greek language were zealously cultivated. Greek law was also important, as it formed the basis of personal and economic relations; and, indeed, the Greeks of the privileged class dominated the political and economic life of the kingdom.

Greek religion accompanied the Greeks to the Near East. Throughout the Seleucid domains, the cult of the Olympian gods coexisted with that of local deities. Greek temples were to be found everywhere—in Phoenicia, Syria, and even in distant Iran. Social contact with the indige-

nous peoples gave the Macedonians and Greeks access to the most diverse religions of the Near East—the cult of Mithra in Iran, Marduk and Ishtar in Babylonia, Astarte and Hadad in Syria and Phoenicia, the Magna Mater (Cybele) in Anatolia, and many others. They assimilated many of these, identifying the foreign gods with their own; but the capacity for assimilation was limited, and the wider Hellenistic civilization spread, the thinner the cover of genuine Greek religiosity became.

The "royal cult"

The kings themselves were accorded diverse honours, again in emulation of Alexander the Great, who had adopted this custom in his Asian territories. The "royal cult" was organized during the second generation of Seleucids by Antiochus I (281–261). The rulers adopted cult names, such as Soter ("the saviour"), Theos ("god"), and Epiphanes ("appearance of god"). This cult was intended to be a unifying influence on the various peoples and nations of the kingdom, but it was mainly official in character and does not seem to have had much popularity.

Because of the prestige of Greek culture, indigenous people often tried to become assimilated into the Greek class, assuming Greek names and participating in the life of the grammar schools. After the 2nd century BC, just when the initial impact of Hellenic culture was declining in the Near East, it was reinforced by the arrival of the Hellenized Romans in the area; thus, Greek culture continued to be a strong influence there after the demise of the Seleucids. (Ed.)

The Seleucids had a double task—to reconcile the Greek settlers with the Iranians, who had previously been masters in the land, and to defend the eastern frontier against a substantial power in Chorasmia (modern Turkistan). By 303 Seleucus I (ruled 312–281) had lost Gandhāra, eastern Arachosia, and Gedrosia to the Mauryan Candragupta; and, although Antiochus I, as co-ruler, had special responsibilities in the east, after the battle of Ipsus (301) western pressures constantly diverted the Seleucids from the east. Later dynastic conflicts and problems raised by Egypt and Pergamum prevented a consistent eastern policy; this led to Iranian irredentism and Greek disaffection.

Seleucid clashes with Egypt

The earlier Seleucid clashes with Egypt are badly documented. On Seleucus I's death (281), Ptolemy II fomented a revolt near Apamea but made peace in 279; Antiochus I (ruled 281–261) was then occupied against the Galatians introduced by Nicomedes of Bithynia, until in 275–274 he defeated them and settled them in the later Galatia. The First Syrian War (274–271) brought Ptolemy II some success, and Antiochus I had to accept the independence of Pergamum. The Second Syrian War (*c.* 260–255 or 253) brought Ionia and many lost territories to Antiochus II (ruled 261–246); it ended in his marriage to Ptolemy II's daughter Berenice and led to a Seleucid revival. But on Antiochus II's death, Ptolemy III (ruled 246–221) championed his sister Berenice, Antiochus' widow, and her young son against Antiochus' former wife, Laodice, and her son, Seleucus II (ruled 246–225). In the Third Syrian (or Laodicean) War, Ptolemy III crossed the Euphrates but without lasting effect. About this time the people later known as Parthians seized Parthyene, and Bactria seceded. Seleucus II was distracted by civil war with his brother Hierax, but he made a compromise peace (236) and soon invaded the eastern provinces, with small result. His eldest son, Seleucus III (ruled 225–223), was fully occupied against Attalus I of Pergamum. He was assassinated in 223. His brother Antiochus III (ruled 223–187) launched the Fourth Syrian War (219–216) against Ptolemy IV. Antiochus lost the Battle of Raphia (217), but an Egyptian rising that followed permanently weakened the Ptolemaic kingdom. Although between 212 and 205 Antiochus III marched as far eastward as Paropamisadae (see below *Roman entrance into Hellenistic affairs*), by the end of the 3rd century the Seleucids were more directly involved in the west than at any time since Seleucus I.

The Greeks in Bactria and India. About 239 BC Diodotus, satrap of Bactria, took advantage of Seleucid weakness to set up an independent Greco-Macedonian kingdom in Bactria, Margiana, and Sogdiana, supported by Iranians who feared a nomad invasion. His son was overthrown by Euthydemus, whom Antiochus III recognized as king after a siege of Bactra (208–206); and Euthydemus' son Demetrius annexed Aria, Arachosia, and perhaps Gedrosia and Carmania. Sometime after 187 another Demetrius of Bactria reached Gandhāra (Punjab); and in about 171 King Eucratides united all the Greek areas around Bactria. Eucratides' son Heliocles was defeated by the nomadic Yüeh-Chih around 135; but there were Greek princes on the Indus and farther eastward much later, the most famous in Greek and Indian tradition being Menander (others are known only from coins). Though inadequately documented, the story of the Greeks in the Far East exemplifies the vitality of the Hellenic tradition.

Pergamum, the smaller Anatolian kingdoms, and Rhodes. Pergamum, a stronghold on the Caïcus River in northwestern Asia Minor, was entrusted by Lysimachus to Philetaerus, who maintained good relations with the Greek world. He bequeathed Pergamum to his nephew Eumenes, who defeated Antiochus I near Sardis and declared himself independent. Eumenes' nephew, Attalus I, defeated Antiochus Hierax's Gaulish allies (*c.* 238) and took the title of *sotēr* ("saviour") and, shortly afterward, that of king. In 133 the last Attalid, Attalus III, left the kingdom to Rome; at its height, through Roman favour, Pergamum had come to embrace much of Asia Minor.

The Attalids of Pergamum

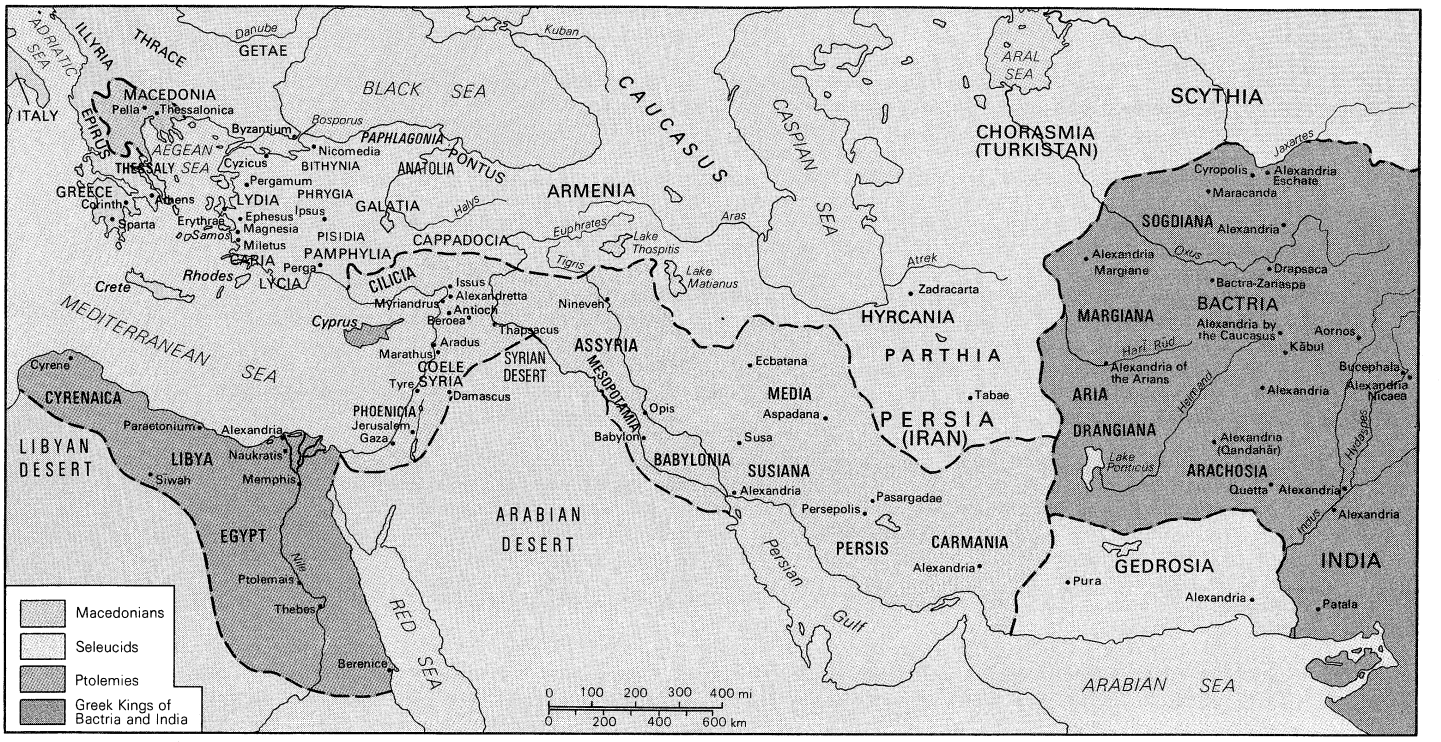
Like the Ptolemies, the Attalids zealously pursued personal gain. Their lands were rich in pitch and timber and they manufactured parchment from hides imported from the Black Sea area. They founded many cities, especially in the newer territories, and even more military colonies for their mercenaries; the cleruchs (soldiers who settled on land allotments) paid a 10 percent tax on produce, but their other subjects paid a fixed amount of grain and a heavy burden of other taxes. Pergamum itself became a splendid city with outstanding amenities and public-health provisions. The Attalids patronized learning, and their library was second only to that of Alexandria. Agricultural experiments were fostered, and industry (resting more than elsewhere upon slave labour) was developed. The Attalids posed as democratic rulers, especially in their capital; but they were widely detested by the Greeks, who regarded them as instruments of Rome.

Bithynia, astride the lower Sangarius, had eluded conquest by Alexander. In 298 or 297 its prince, Zipoetes, declared himself king. His successor Nicomedes I joined the Greek cities against Antiochus I but later brought the Gauls (Galatians) over into Asia Minor as mercenaries; in about 260 he founded Nicomedia. The later kings cultivated Greek friendship by generous gifts and founded many Greek cities, mainly on sites already inhabited. The population, largely Thracian, was substantially Hellenized. Of the later kings, Prusias I supported Rome against Antiochus III, and Prusias II played an ambiguous role in the Third Macedonian War; after a career as a noteworthy flatterer of Rome, the latter fell victim to a dynastic imbroglio (149).

Pontus, on the southern shore of the Black Sea, east of the Halys, was seized by Mithradates, the nephew of a tyrant of Cios, in 302; he made himself king and cultivated friendly relations with the Seleucids. To the south, Cappadocia, stretching from Lycaonia to the Euphrates and as far as the Taurus Range, had also remained independent under its satrap, Ariarathes; about 255 Ariarathes III became king and cemented good relations with Seleucus II. Both Pontus and Cappadocia remained feudal and mainly Iranian in outlook and custom.

The Galatians, intruders in the Greek world, were a branch of the same Celtic peoples who had invaded Italy in the early 4th century. Following their invasion of Macedonia and their attack on Delphi, a group including the Tectosages, Trocmi, and Tolistobogii (Tolistoages) was brought into Asia Minor by Nicomedes I of Bithynia in 277 and attacked Cyzicus, Erythrae, and Didyma. Settled by Antiochus I around the upper Sangarius and middle Halys, they were feared and hated by the Greeks. (Attalus I's victories over them were later commemorated in the frieze of the great altar of Zeus at Pergamum.) The Galatians kept their language, customs, and tribal organization down to the Christian era.

Its skilled navy and commercial strength enabled the



The kingdoms of Alexander's successors *c.* 185 BC.
From J. Thomson, *Everyman's Classical Atlas*

island city of Rhodes to play an independent vigorous political role, especially in defense of the freedom of the seas. In 220 it fought Byzantium to keep the straits open, and it later made war on the Cretan pirates. In 305–304 the Rhodians successfully withstood the siege of Demetrius Poliorcetes. About 250 they defeated Ptolemy II off Ephesus in circumstances that remain obscure; generally, however, Rhodes was closely aligned with Egypt and inherited its control of the Aegean. Donations from other states to help repair the damage caused by a serious earthquake of *c.* 227 indicated Rhodes's high prestige throughout the Greek world. The richer families held power in a limited democracy and used their wealth to stave off social troubles; the 1,000,000 drachmas raised annually from import and export duties in 170 are an indication of Rhodes's wealth. After 167 Rhodes fell foul of its friends in Rome and suffered greatly in consequence. The Rhodian schools of philosophy and rhetoric were famous, and the city possessed many works of art; its maritime law may be partly incorporated in the Byzantine code known as the Rhodian Sea Law.

The Antigonids' "heroic monarchy"

Macedonia: government and foreign policy. The Antigonids retained some of the traits of "heroic monarchy" evident under the Argeads. The king was the state; but the army assembly elected a regent and tried cases of treason. The monarchy was neither absolute nor personal; the king was first among his peers, neither a conqueror nor (to the Macedonians) a god. Macedonia was a poor land: the land tax raised only 200 talents annually. It was a country of peasant landowners (though the kings granted conquered land to settlers and veterans); mines and forests were royal possessions. The capital, Pella, became a Hellenized city and nominally autonomous. The upper classes spoke Greek and worshipped Greek gods. Among the few new cities were Thessalonica and Cassandrea. Antigonid power, solidified by Antigonus II Gonatas (ruled 276–239), rested on the citizen army; Macedonia avoided the problems of mercenary armies familiar to the Ptolemies and Seleucids.

Not until the death of Pyrrhus of Epirus (272)—who had almost conquered Macedonia in 274—did Gonatas feel secure. For 50 years Macedonia tried to prevent any strong power from arising in Greece. To that end Gonatas maintained garrisons in Corinth, Demetrias, Chalcis, and Eretria; friendly governments in Argos, Megalopolis, Elis, and Megara; and a tacit agreement with Aetolia. A coalition raised by Athens and Sparta and backed by Ptolemy II took four years to suppress (266–262) in the Chremonidean War, named after Chremonides, who introduced the war measure at Athens; Gonatas' defeat of the Egyptian fleet off Cos may have occurred in 262. Ten years later, Alexander, governor of Corinth, revolted; and, although Gonatas recovered the city on Alexander's death, he soon lost it again to the Achaeans. Both he and his successors—Demetrius II, Antigonus III, Philip V, and Perseus—were much concerned with repelling the northern barbarians. But in Greece the policy of those rulers turned on their relations with the Achaean and Aetolian confederations, which were to involve Macedonia with the growing power of Rome (see below *The Greek leagues and the Antigonids in the 3rd century BC*).

The Greek states. The Hellenistic Age saw a number of changes affecting the Greeks, both on the mainland and overseas.

The particularism of the *polis* underwent modification. In the mainland cities the machinery of government still functioned; but many were controlled by Macedonia, and relations with Macedonia became the main political issue alongside the rival interests of rich and poor. Elsewhere much of the content had gone out of political life. There was considerable material prosperity; the economic centre had shifted to Asia (especially to Rhodes), but Corinth, Delos, Pagasae, and Ambracia all flourished. In many ways life had improved. Attempts were made to humanize the practice of war. Many cities gained neutral status as holy places; and inscriptions record grants of immunity from the widespread custom of reprisals. Arbitration was common, and cities frequently invited outside commissions to settle lawsuits.

Freeing of slaves was frequent, some slaves buying their own freedom. Greater wealth led to new and more elaborate festivals and to an increase in social clubs; in some areas, like Boeotia, excessive social life became an abuse. Though depopulation was not serious before the 2nd century, some family limitation was practiced, probably through infanticide or abortion. In the 2nd century, however, the Roman wars brought a deterioration to most Greek cities and utter ruin to some.

In the 3rd century federalism increased, thus providing compensation for the weakness of the separate cities. The Boeotian and Arcadian leagues were both important,

but the most effective leagues were those in Achaea and Aetolia.

The Aetolian and Achaean leagues

The Aetolian cantons were federally organized by the early 4th century, and after 300 the league included most of central Greece. Outlying areas enjoyed a loose form of citizen exchange (*isopoliteia*), and both Delphi and the Delphic amphictyony (association of neighbouring states) were under Aetolian control. The league had a representative council and a primary assembly meeting in spring and (at Thermum) in autumn. A cavalry leader, secretary, and other magistrates assisted the general, who was military leader; but a board of 30 *apoklētoi* —a council committee under the general—directed policy. The land was poor; this encouraged mercenary service abroad and piratical expeditions, especially as the accessible seaboard grew. In the First Macedonian War the Aetolians fought with Rome against Macedonia.

An Achaean league existed in the 5th century, but it had lapsed. In 280 Dyme, Pharae, Patras, and Tritaea reformed it; and by 272 the league of ten cities, including Aegium with the federal centre of Zeus Homarios, was complete. The accession of Dorian Sicyon under Aratus in 251 introduced an anti-Macedonian policy, which lasted until the rise of Cleomenes of Sparta (see below *The Greek leagues and the Antigonids in the 3rd century BC*). Achaea, like Aetolia, had a primary assembly and a representative council. In 255 its two generals gave place to one, who might not serve in consecutive years; there were subsidiary magistrates, including a cavalry commander and secretary, and a ten-member board meeting under the general. The organization of the assemblies is still controversial; but apparently during the 3rd century the assembly met four times a year at Aegium, together with the council and magistrates, for routine business and elections. These meetings (*synodoi*) were restricted in the 2nd century to the council and magistrates. Special meetings (*synklētoi*) could always be summoned for special business, and after 200 they became mandatory to decide peace or alliance or to hear communications from the Roman Senate. There were federal laws and courts (as well as those of the cities) and, in the 2nd century, federal coins.

The Achaean League eventually embraced the whole Peloponnese. Its failure to strike a satisfactory compromise in its relations with Rome and an overweighting in favour of the propertied classes, who were unable or unwilling to make concessions to the increasingly radical poor within the cities, led to the league's downfall; but this does not detract from what it achieved during the 3rd and early 2nd centuries.

Aspirations of the large cities

Such cities as Athens, Thebes, and Sparta found it hard to accept a diminution in scale and opportunity in the Hellenistic Age. For 60 years after Alexander's death, Athens aspired to an independent policy despite its defeat in the Lamian War (322). It had 10 years of conservative administration under Demetrius of Phaleron, but after its "liberation" (307) it indulged in wild adulation of Antigonus I and Demetrius I. The Athenian and Spartan revolt against Antigonus II (the Chremonidean War) was crushed in 262; subsequently Macedonia controlled and garrisoned Athens until 229, when the governor Diogenes sold the city its freedom. Thenceforth it pursued a modest role under a conservative democracy, enjoying Ptolemaic protection. Under both the Antigonids and the restored democracy, Athens enjoyed high repute because of its past and even more as the home of the philosophical schools and the New Comedy.

Thebes, destroyed by Alexander, was restored in 316 by Cassander; but its position on the fringe of Aetolian- and Macedonian-held areas restricted its political life. A supple policy, mainly based on friendship with Macedonia, kept it independent; but the Aetolians broke its military power in 248, and internally it suffered from inadequate administration, social conflict, and extravagant expenditure, foreshadowing its early 2nd-century clashes.

Sparta, consistently anti-Macedonian, warded off both Demetrius I and Pyrrhus. King Areus joined in the Chremonidean War but fell at Corinth; and under Agis IV (ruled 244–241) Sparta was briefly allied with Achaea against Aetolia. Agis was the first to attempt to deal with social ills found elsewhere but most evident at Sparta, where wealth was in the hands of the few and where a citizen body that depended on the possession of land had dwindled to a few hundred. Agis cancelled debts but was murdered before he could divide up the land as he proposed. His program was taken over and put into effect by Cleomenes III (ruled 235–222), whose imperial ambitions led the Achaean Aratus to call in Antigonus III. After Cleomenes was defeated at Sellasia and fled to Egypt, Sparta became a constant source of friction with Achaea. Revolutionary policies were revived under Machanidas and Nabis but later they degenerated into faction; and Sparta declined into a sterile imitation of its Lycurgan past, becoming a tourist centre under the Roman Empire.

The western Greeks

By the middle of the 3rd century BC, the western Greeks were drawn irrevocably into the current of Roman history. In 317, after much faction, Agathocles, a man from Thermae, made himself tyrant at Syracuse. Engaged in a Carthaginian war, which was going poorly, he slipped out of Syracuse to land in Africa; there Ophellas of Cyrene joined him but was soon murdered by Agathocles. The war was inconclusive, but Agathocles on returning to Syracuse made a peace that established the Halycus River as the boundary with Carthage in Sicily. In 305 Agathocles took the title of king; he later campaigned in Italy and gained Corcyra, which, however, he gave to Pyrrhus of Epirus as dowry with his daughter Lanassa. Dissension in his family over succession caused Agathocles on his death in 289 to restore freedom to Syracuse. His reign had wasted Sicily by inconsequential policies and savage massacres.

Pyrrhus, the Molossian king of Epirus (ruled *c.* 307–272), came nearer than anyone to bringing southern Italy and Sicily within the Hellenistic sphere. A relative of Alexander, he was perhaps the stormiest figure of his generation. Expelled from Epirus in 302, he was restored by Ptolemy I and Agathocles and set about giving a Hellenistic stamp to the monarchy of Epirus. His aim was Macedonia; and he seized part of it, with Thessaly, in 288 but was driven out by Lysimachus in 285. In 280 Tarentum invoked Pyrrhus' help against Rome; but, after victories at Heraclea (280) and Asculum (279), he crossed to Sicily, where anarchy had followed Agathocles' death. He almost drove the Carthaginians from the island; friction with the Greeks, however, led to his return to Italy and, after a drawn battle at Beneventum, to Greece. Failing in attacks on Antigonus II and Sparta, he perished in a night melee in Argos (272). Pyrrhus' policy was opportunist, and his real aims, if any, remain unfathomable; his return to Greece once more, however, severed the threads linking East and West.

At Syracuse Hieron II (ruled *c.* 270–215) seized power and took the royal title after defeating the Mamertini, Campanian mercenaries of Agathocles, who had seized

From W. Shepherd, *Historical Atlas*



The Aetolian and Achaean leagues.

Messana (Messina). The Mamertini invoked Roman aid. Hieron at first allied himself with Carthage; but in 263, after a defeat, he made a peace that initiated a period of some 50 years in which he ruled as a successful client prince of Rome. Thenceforth southern Italy and Sicily remained part of the Roman world.

Relations among the Hellenistic states, 275 to 30 BC. *Ptolemaic policy in the Aegean and Asia Minor.* Together with Coele Syria, the Ptolemaic overseas possessions were valuable for their wealth and formed an empire for the defense of Egypt; under Ptolemy IV (ruled 221–*c.* 204) and Ptolemy V (ruled *c.* 204–180), the possessions were mostly lost to Syria, or they either asserted or were granted independence. By 195 Egypt held only Thera and a few Cretan towns, having also lost Coele Syria, the main object of the long series of 3rd- and 2nd-century wars between Syria and Egypt.

The Syrian–Egyptian wars

Of these wars the first four (see above *The Seleucid kingdom*) left Coele Syria Egyptian; but the weakness of Ptolemy IV after the Battle of Raphia (217) and Antiochus III's experience and prestige made Syrian retaliation inevitable. By 203 Antiochus was interfering in Caria; his secret agreement with Philip V of Macedonia in 203–202 to dismember the Ptolemaic possessions after the accession of the boy king Ptolemy V probably assigned Caria to Philip. After advancing to Gaza and retreating to the sources of the Jordan, Antiochus won a victory at Panium (200), thereby gaining Coele Syria, which thenceforth was Selucid and under a general of Coele Syria and Phoenicia. In 195, having tried unsuccessfully to annex Cyprus, Antiochus made peace, marrying his daughter Cleopatra to Ptolemy V. Peace lasted 25 years, during which (in 186) Ptolemy V at last suppressed rebels who had held out in Upper Egypt since early in the previous reign. Antiochus, following on a defeat by Rome (see below *Roman entrance into Hellenistic affairs*), lost his western provinces; he was succeeded by his son Seleucus IV in 187. Seleucus was assassinated in 175 and succeeded by his brother Antiochus IV Epiphanes, who ruled until 163.

Antiochus had seized the throne with Pergamene help and did much to restore the kingdom. In 170 the regents of Ptolemy VI Philometor (ruled 180–145), the younger son of Ptolemy V and Cleopatra, declared war on Antiochus, hoping to recover Coele Syria. Both sides appealed to Rome; but the Senate, by then at war with Perseus of Macedonia (see below *Roman conquest of Macedonia and Greece and the acquisition of Pergamum*), delayed a decision. In 169 Antiochus invaded Egypt, besieged Alexandria, then returned home, leaving Ptolemy VI in control; Ptolemy, however, joined his sister Cleopatra II and his brother Ptolemy VIII against Antiochus and once again appealed to Rome.

The next year Antiochus seized Cyprus and again invaded Egypt. By then, however, Rome had defeated Perseus, and it ordered Antiochus to evacuate both Egypt and Cyprus. Antiochus yielded. The incident marked a decisive stage in Syrian decline; Egypt, too, was shaken, and thenceforth it could pursue an independent policy only through Roman indifference.

Syria was further weakened by a revolt of the Maccabees against the Hellenizing Judaean aristocracy, a movement that eventually ended in the establishment of a Jewish kingdom. In 150 Ptolemy VI married his daughter to the Syrian pretender, Alexander Balas; and another son of Seleucus IV, Demetrius I (ruled 162–150), was defeated and killed by Alexander Balas. Three years later Demetrius' son, Demetrius II (ruled 145–139; 129–125), landed in Cilicia to challenge Balas, whereupon Ptolemy transferred support to him in the hope of receiving Coele Syria; Demetrius defeated Balas in the Battle of Oenoparas (145), and Coele Syria remained Seleucid. In 132 Ptolemy VIII successfully raised a usurper against Demetrius II, who was tortured to death in 126 or 125; later, however, Ptolemy supported Demetrius' son, Antiochus VIII Grypus.

Dissolution of Seleucid Syria

The years 123–83 witnessed the final dissolution of Seleucid Syria, rent with dynastic strife and prey to Jews and Arabs within and to Parthians without. Tigranes of Armenia seized the remnants of the Seleucid kingdom in 83, which by then was ripe for Roman annexation. Ptolemaic Egypt survived semi-independent for another half century.

The Greek leagues and the Antigonids in the 3rd century BC. Soon after bringing Sicyon into the Achaean League, Aratus in a peacetime coup seized the Acrocorinth. Antigonus II Gonatas made no attempt to recover this position in southern Greece, but he leaned more heavily on the friendly tyrants in the Peloponnesian cities. In 241 Achaea made peace with Macedonia (and also with the Aetolians, who had recently expanded into central Greece and had been raiding the Peloponnese); nevertheless, Aratus launched a series of attacks on Argos and Athens.

Upon succeeding his father Antigonus, Demetrius II (ruled 239–229) made an alliance with Epirus, marrying the princess Phthia (Philip V's mother). This development provoked an Achaeo-Aetolian alliance, and, in 239–238, war broke out between Macedonia and the leagues. The Aetolians coveted Epirote Acarnania, and Aratus aimed at bringing all the Peloponnese under Achaea. Aratus seized Cleonae and Heraea, but he failed at Argos, where in 235 Aristomachus succeeded his brother Aristippus as tyrant. That same year, however, Lydiades resigned his tyranny at Megalopolis, which joined the Achaean League, as did Orchomenus and Mantineia. Demetrius, meanwhile, was busy fighting the Aetolians; in 237–236 he entered Boeotia and took Megara, thus driving a wedge between the leagues. In 229 Demetrius died while fighting in the north. His son, Philip (later Philip V), was only nine.

Macedonia was imperilled. The Aetolians seized Phthiotis, Thessaliotis, and Hestiaeotis; but Antigonus Doson, a nephew of Gonatas, was elected regent by the army and restored the position. In three years Antigonus Doson reestablished the northern frontiers, recovered most of the Aetolian conquests, and seized part of Phocis. Doson was recognized as king (Antigonus III), but he loyally held the position in trust for Philip. In 227 he made a naval expedition to Caria, an obscure episode. Meanwhile, the monarchy in Epirus had been swept away, and Illyrian attacks caused the state to disintegrate. Ambracia and Amphilochia went with Aetolia, Athamania became independent, and Epirus itself eventually joined the Illyrians. Farther south, Demetrius' death led to the freeing of Athens (229–228), and Aristomachus surrendered his tyranny to join Achaea (228). Under Argive and Megalopolitan influence, Achaea became more hostile to Sparta, where the revolutionary Cleomenes III had succeeded in 235. Shortly before, perhaps as compensation for its greater burden in the war, Achaea had ceded Tegea, Mantineia, Caphyae, and Orchomenus to the Aetolians; in 229 they were given to Cleomenes.

Cleomenes' invasion of Megalopolitan territory initiated the Cleomenean War with Achaea (229–222), which quickly developed in his favour. In 227 he deposed the Spartan ephors, raised the citizen body to 4,000 by radical land redistribution, and introduced Macedonian tactics. Aratus, seeing the league dissolve and fearing social revolution, appealed to the Macedonian Antigonus Doson, agreeing in 225–224 to surrender Corinth in return for Macedonian help. With the Macedonian army confronting Cleomenes at the Isthmus, Argos revolted from Cleomenes and forced him to withdraw into Laconia. In 223 Cleomenes' Arcadian gains fell away; and, though in winter (223–222) he seized and destroyed Megalopolis, his paymaster Ptolemy (who had abandoned Achaea a little before) deserted him. Defeated at Sellasia (222), Cleomenes fled to Egypt, where he perished in a rising shortly afterward. Antigonus occupied Sparta, but a Dardanian invasion soon drew him back to the north; after a victory there, he died of consumption in 221.

The Hellenic Alliance

In 224 Antigonus had created a new organization to control Greece—the Hellenic Alliance, a loose body of leagues (not cities), including Achaea, Epirus, Phocis, Boeotia, Acarnania, Opuntian Locris, Euboea, Thessaly, and (nominally) Macedonia. The Macedonian king was president; a council made all decisions, which each state had to ratify. This alliance threatened Aetolia with encirclement, and in 220 the Achaeans clashed with Aetolians operating in Messenia. The ensuing Social War was essentially a conflict between Achaea and Aetolia, with Antigonus' successor Philip V (ruled 221–179) and the alliance play-

ing reluctant parts. Elis and Sparta joined Aetolia, and various Cretan cities aligned themselves on the two sides for domestic reasons. An impressive list of demands was presented to Aetolia. In the three years of fighting Philip revealed tactical skill and personal energy—crushing internal opposition among his ministers and sacking Thermum. But when peace was concluded at Naupactus (217), both sides kept what they held.

Already Philip's eyes were elsewhere. In 221 the Illyrian chieftains, Demetrius of Pharos and Scerdilaidas, had broken the terms imposed by Rome in 228 by sailing southward on a piratical foray. In 219 a speedy policing expedition (the Second Illyrian War) had crushed Demetrius, who, as an exile at Philip's court, now urged him to pursue Macedonian (and Illyrian) interests in the Adriatic.

Roman entrance into Hellenistic affairs. Through the Hellenic Alliance, most of Greece was drawn into a new conflict between Macedonia and Rome. At first Philip enjoyed mixed fortunes in Illyria: in 215 he approached Hannibal of Carthage, and a treaty was made in very general terms to secure Philip's Illyrian aims. In 216–212 Philip suffered a naval setback but reached the Adriatic coast by land; in Messenia his inept policy alienated Achaea. In 211 the Romans engineered an Aetolian alliance, under which the Aetolians were to have conquered territory while the plunder went to the Romans or was shared. Soon Elis, Sparta, and Messenia came in on the Aetolian side, along with Attalus of Pergamum (to whom the Aetolians sold Aegina, taken from Achaea). In 209 and 208 Philip was characteristically active defending his allies, and Attalus retired to Asia to meet the attack of Prusias of Bithynia. Neutral attempts at mediation—by Egypt, Rhodes, Chios, and Athens—were fruitless. In 207 and 206 Roman inactivity allowed Philip to invade Aetolia and force it to a separate peace; meanwhile, the Achaean army, reorganized by Philopoemen, had destroyed the Spartan tyrant Machanidas at Mantineia (207). The Peace of Phoenice, made with Rome in 205, represented a limited success for Philip, who then turned his attention to the Aegean.

Simultaneously Antiochus III was looking in the same direction. After Raphia (217) he had spent three years crushing the revolt of his cousin Achaeus and recovering some of his territory in Asia Minor. He then led his forces eastward, covering Media, Parthia, Bactria, Paropamisadae, and Arachosia and ending at Gerrha in Arabia. Antiochus thereby won wealth and the title of great and reasserted Seleucid authority in Media. His treaties with rulers farther eastward, however, merely disguised their virtual independence. In 204–203 he took Amyzon in Caria; but Ptolemy IV's death and the accession of a boy, Ptolemy V Epiphanes, changed the whole situation.

In winter 203–202 Philip, who had already been treacherously attacking Rhodes and raising money by piracy, compacted with Antiochus to divide the overseas possessions of Ptolemy; and in 200 Antiochus recovered Coele Syria (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 202 Philip campaigned in Thrace, seizing the Aetolian possessions of Lysimacheia, Chalcedon, and Cios; in 201 he sailed a new fleet across the Aegean, attacking Rhodes and Pergamum and campaigning in Caria.

Roman declaration of war against Philip

Attalus and Rhodes appealed to Rome, and in 200 a Roman embassy declared war on Philip as he was besieging Abydos. Roman motives have been variously assessed: Philhellenism, fear of Philip, fear of Philip and Antiochus together (in the light of exaggerated accounts of their pact), and pure imperialism have all been suggested. Probably motives were mixed; but irrational alarm undoubtedly figured prominently among them.

Though isolated, Philip resisted for three years (200–197). In 198 the Roman Titus Quinctius Flamininus defeated him on the Aous in Illyria, forcing him back into Macedonia. Aetolia and Achaea (with some opposition from Argos) joined the Romans. There were negotiations in winter 198–197, but, once assured that his command would be renewed, Flamininus made sure they collapsed. In June 197 Philip was decisively beaten at Cynoscephalae in Thessaly, the Aetolians contributing greatly to the victory.

The Roman demands had been gradually stepped up until they included the general requirement that all Greeks be free; the proclamation of this policy at the Isthmian Games of 196 aroused tremendous excitement. Philip was restricted to Macedonia; much of Thessaly was organized in federations; Aetolia received some territories but emerged discontented. The Greeks were freed; but the Romans expected their willing collaboration (failure to appreciate this was later to cause disaster). Nabis, the tyrant of Sparta, was obliged to free Argos (made over to him by Philip) but kept his throne. In 194 the legions left Greece.

Antiochus, meanwhile, had reached the Hellespont (the Rhodians having withdrawn opposition upon the news of Cynoscephalae). Smyrna and Lampsacus appealed to Rome, and Antiochus was ordered to withdraw. But when at a conference at Lysimacheia (summer 196) the Romans bade him refrain from attacking Greek cities or Ptolemy's possessions, Antiochus parried this with a pertinent reference to the Greeks in Italy and announced his forthcoming alliance with Ptolemy (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 195 Hannibal fled to Antiochus' court; and in 193, through a Syrian embassy at Rome, Antiochus was ordered to stay out of Europe or see the Romans intervene in Asia. Probably neither side sought war. But in 192 the Aetolians tried to raise a coalition against Rome, and, though Philip wisely declined and Nabis was soon assassinated, Antiochus felt he must seize the opportunity; in autumn 192 he crossed, inadequately prepared, to Greece.

Elis, Boeotia, and Amynander of Athamania joined Antiochus; but Achaea and Philip both fought alongside Rome. In 191 Antiochus was defeated at Thermopylae and returned to Asia. The Romans granted the Aetolians a truce, which was renewed in 190 when Lucius Scipio as consul, with his brother Africanus as his legatus, crossed to Greece and marched northward through Macedonia to the Hellespont. Three naval victories—at Corycus (191) and at Side and Myonnesus (190)—gave Rome's allies, including Rhodes and Pergamum, command of the seas. Antiochus evacuated Thrace; and in winter 190–189 his army was defeated by the smaller but more disciplined Roman force at Magnesia-by-Sipylus (Magnesia ad Sipylum).

In 189 the Aetolians were given terms—the first recorded example of a treaty requiring the other party to "preserve the majesty of Rome." In Asia the new consul, Manlius Vulso, turned a nominally punitive expedition against the Galatians into a large-scale plundering of Asia Minor. The Peace of Apamea (188) fixed Antiochus' boundary at the Taurus Range and limited his fleet and armaments. Excluded from Asia Minor (except eastern Cilicia), his realm was thenceforth firmly centred in Syria. The program of liberation was necessarily compromised. Cities that had been free or had abandoned Antiochus before Magnesia remained free; the rest, along with large parts of Asia Minor (excluding Lycia, Caria, and the Maeander Valley, which went to Rhodes), were assigned to Eumenes II of Pergamum, the main beneficiary of the peace.

Roman conquest of Macedonia and Greece and the acquisition of Pergamum. Despite Philip's discontent with the outcome of the peace, his reorganization of Macedonia between 188 and his death in 179 (he built up the royal revenues by taxes, the encouragement of trade, and the exploitation of mines) was not necessarily with a view toward war on Rome, as some sources suggest. Constant complaints to Rome from the Greeks, especially in Pergamum and Thessaly, resulted in decisions invariably to Philip's disadvantage. He was obliged to relinquish Aenus and Maroneia in Thrace, which he had seized. From 184 onward, he campaigned and intrigued in Thrace, but his plan to raise the Bastarnians against the Dardanians fell through upon his death in 179. On succeeding, his son Perseus (ruled 179–168) revised Philip's policy and cultivated friendly relations with the Greek states and the amphictyonic council and made marriage alliances with the Seleucids and with Prusias of Bithynia. The Seleucid alliance was upset, however, in 176–175, when Perseus' enemy Eumenes sponsored Antiochus IV as Seleucus IV's successor (see above *Ptolemaic policy in the Aegean and Asia Minor*). In 172 Eumenes visited Rome and accused Perseus of hostility to Rome; the Romans resolved to elim-

End of the Antigonid monarchy

inate Perseus, perhaps fearing a repetition of the situation of 201 because Antiochus IV was beginning to loom large in the Middle East. The Third Macedonian War (172–168) began in the winter of 172–171; in 171 a Roman force crossed over to Greece. For three years the Romans made little headway. Generally sympathy was with Perseus, and Epirus and the Illyrians under Genthius joined him; both Rhodes and Eumenes himself appear to have been compromised. In June 168, however, Lucius Aemilius Paullus defeated Perseus at Pydna; and the Antigonid monarchy came to an end.

Macedonia was divided into four separate tribute-paying (if nominally independent) republics with capitals at Amphipolis, Thessalonica, Pella, and Heraclea Lyncestis. Athens was rewarded for its loyalty by the gift of Lemnos, Imbros, Scyros, Delos, and Haliartus. But Epirus was savagely crushed, its population enslaved as an example to others.

Greece was disturbed and uneasy after this war. In Achaea from 189 onward there had been constant trouble with Sparta (brought into the league in 188) and Messenia (where a revolt was crushed in 182). Roman indecision and occasional malevolence heightened the general feeling of unease. After Perseus' defeat, pro-Roman politicians like Callicrates gained power, and 1,000 leading citizens were deported to Rome and kept there until 151–150; they included the historian Polybius. Despite this, internal problems continued to arise, and at each step the contenders appealed to Rome.

In Macedonia the new republics were popular only with the richer classes. In 149 a pretender, Andriscus, took the name of Philip and carried Macedonia before him. He destroyed one legion (148) but was defeated and sent to Rome by Metellus. Macedonia was then made a province, and its proconsul was given responsibility also for Illyria. About the same time general political and social discontent precipitated revolt in Achaea. Further trouble arose over Sparta; and Diaeus, a popular Achaean leader, adopted a truculent attitude toward Rome. In 147 Lucius Aurelius Orestes was sent out to Greece and demanded that Sparta and also Corinth, Argos, Orchomenus, and Heraclea Trachinia be made independent of the league. Further envoys found the Achaeans stubborn and bent on war with Sparta. In 146 a general conflict flared up, including Boeotia, Phocis, and Locris; but Metellus defeated the Achaean force in central Greece, and Lucius Mummius sacked Corinth as an act of deliberate punitive terrorism. The Achaean League was temporarily dissolved and, together with those states that had supported it, was put under the control of the proconsul of Macedonia. Rome had now firmly established itself in the Balkans, and Greek political freedom was at an end.

In Asia the war with Perseus had left Eumenes II under a cloud. The Galatians and Prusias II of Bithynia had exploited this fact by attacking him; but he was succeeded by Attalus II (ruled 159–138), his brother, who recovered the goodwill of Rome. Prusias was forced to desist from war in 156, and Pergamum grew in prosperity. Attalus III (ruled 138–133), a son of Eumenes II, succeeded his uncle; and on his premature death for unknown reasons left his kingdom (apart from Pergamum itself and its territory) to the Roman people. Attalus III's death precipitated a civil war because Aristonicus, a bastard son of Eumenes II, claimed the throne. Pergamum and most of the cities supported Rome; Aristonicus, with some following among the Greco-Macedonian population, was mainly supported by the rustic and slave elements. He cancelled debts, freed slaves, and set up a centre called Heliopolis in Mysia, the significance of which has been much debated: he may have been trying to establish some kind of mystical utopia; at least it served as a means to rally supporters.

The disappearance of Pergamum

In 130 the rebels were defeated and Aristonicus was taken prisoner. In 129 Manius Aquilius and a commission settled the fate of the Pergamene territories. Many fringe lands in the interior were given to such kingdoms as Pontus and Cappadocia. The Thracian Chersonese and Aegina were placed under the proconsul of Macedonia. But the heart of the kingdom—Mysia and the Troad, Lydia, southwest Phrygia, and perhaps parts of Caria—became the Roman province of Asia. The disappearance of Pergamum just as the Seleucid kingdom lost Mesopotamia to the Parthians under Mithradates I (who had taken Demetrius II prisoner in 140–139) marked a further stage in the disintegration of the Hellenistic world and the transformation of the Greeks from allies to Roman provincials.

Rome and Mithradates VI. The kingdom of Pontus had gained some prominence early in the 2nd century, when Pharnaces had attempted to expand but had been forced to surrender most of his gains in a peace underwritten by Rome (182–179). Mithradates V Euergetes (ruled *c.* 150–*c.* 120) was given part of Phrygia for helping the Romans against Aristonicus, but upon Mithradates' assassination the Romans revoked the gift—a source of bitterness to his successor, Mithradates VI Eupator (died 63 BC), who became effectively king of Pontus in 112. After gaining the realm of Bosporus, north of the Euxine, and expanding his territories eastward to Colchis and Trapezus, Mithradates VI agreed with Nicomedes III of Bithynia to attack Paphlagonia and Cappadocia. Rome intervened, compelled the evacuation of both countries, and placed Ariobarzanes on the Cappadocian throne (95). After being expelled and restored by Sulla (92), Ariobarzanes soon became a refugee at Rome along with Nicomedes IV of Bithynia, whom Mithradates then replaced with a puppet. A Roman commission restored both. Later, when the Romans provoked an invasion of Pontus, Mithradates decided on war (88). Asia was seething with hatred toward Rome because of its oppressive system of tax collection, both in the cities and the countryside. Mithradates was therefore welcomed everywhere. Three Roman armies were overthrown, and from Ephesus Mithradates organized a massacre of 80,000 Italians. His fleet crossed the Aegean to Athens, which had joined him; in Asia, only Rhodes resisted him.

The Roman general Sulla recovered Athens in 87–86, and, after two Roman victories at Chaeronea and Orchomenus, the war in Europe was over (86). A second Roman army, sent out by the supporters of Sulla's rival, Marius, was successful in Bithynia; and in 85 Sulla made an agreement of doubtful validity with the Pontic general Archelaus—the Peace of Dardanus, by which Mithradates surrendered all his gains since 88. Nicomedes and Ariobarzanes were restored, and collaborating cities in Asia were subjected; Rhodes was rewarded with additional territory. Effectively, this was the end of any freedom for the Greek cities of Asia.

Between 84 and 74 Mithradates was subject to some provocation from Roman officers; after Sulla's death the Romans broke off negotiations with him. They were increasingly concerned, however, with the problem of piracy in the Mediterranean; its centre in Cilicia Trachea was excluded from the province of Cilicia, which was set up in 80. Hostilities with Mithradates broke out again in 73, when he invaded Bithynia. Between 73 and 70 Lucius Lucullus defeated him, took over Pontus, and carried out a humane reorganization of Asia (thereby antagonizing financial circles at Rome). In 69–68 Lucullus advanced into Armenia, where Tigranes had given asylum to Mithradates; but his army refused to go farther, and by 67 Mithradates was back in Pontus, and Lucullus was without a command.

The arrival of Pompey

Intrigues at Rome had brought about the appointment of Gnaeus Pompeius (Pompey), first, to act against the pirates—which he did in 67—and then to take over the Mithradatic War with powers far greater than those of Lucullus. In 66 Pompey marched from Cilicia, where he had ended the piratical war, over Taurus to occupy Cappadocia; and by the end of the summer he held Pontus. Mithradates fled to Colchis and later to the Crimea, where he committed suicide in 63. Pompey meanwhile settled Asia Minor, now consisting of four Roman provinces—Asia, Cilicia, Bithynia, and Pontus—and a number of client princes, notably in Cappadocia, Galatia, and Paphlagonia. In 66–65 Tigranes submitted, and Armenia became a client state. After campaigning in the Caucasus, Pompey in 64 entered Syria, which was made a province; evidently the decision had already been taken at Rome. A last Seleucid, Antiochus XIII, briefly restored by Lucullus, was deposed and murdered by the Arab shaykh Sampsiceramus of Emesa.

In Judaea, where the high priesthood had developed royal characteristics and civil war over succession had been in progress since 76, Pompey gave his support to John Hyrcanus II and expelled his younger and more vigorous brother, Aristobulus; Judaea, however, became tributary, and Hyrcanus was styled ethnarch. Pompey's settlement in this region was on similar lines to that in Anatolia. Commagene took its place beside Armenia, Galatia, and Cappadocia as part of the cordon of client kingdoms between Roman territory and Parthia. Tarcondimotus of Kastabala in Amanus covered the communications between the provinces of Cilicia and Syria; and Judaea was a buffer to the south, with the Ituraean and Nabataean kingdoms flanking it.

This settlement—the first to be carried out by a Roman general without a senatorial commission—brought Pompey great prestige and a vast patronage, to which he was later to turn in his war with Caesar. It also marked the real end of the Hellenistic order. The contrast was thenceforth between a Mediterranean world, under Rome, and the Iranian East including Mesopotamia. The one exception was the still nominally independent kingdom of the Ptolemies in Egypt.

Roman annexation of Egypt. Upon the death of Ptolemy VIII (Euergetes II) in 116, the kingdom was divided between his three sons—Ptolemy IX Philometor (Soter II), who took Egypt; Ptolemy X Alexander, who took Cyprus; and a bastard, Ptolemy Apion, who took Cyrene. In 96 Apion died, leaving Cyrene to Rome (but it was made into a quaestorian province only in 74); and in 88 Ptolemy IX murdered Ptolemy X and reunited Egypt and Cyprus. During the Mithradatic War he pursued a cautious neutral policy, in view of the fact that Mithradates had gained possession of his two sons, who were in Cos when the Pontic king seized it. On the death of Ptolemy IX in 81, Sulla sent the son of Ptolemy X (who had escaped to Rome) to succeed him; as Ptolemy XI his reign was short (19 days) and important mainly for his will authorizing the Romans to annex Egypt and Cyprus when they wished.

Ptolemy IX's sons, however, soon appeared at Alexandria. The elder, Ptolemy XII—usually known as Auletes, the Fluteplayer—spent 20 years and vast sums of money in bribes to get his position sanctioned at Rome. Thenceforth the main importance of Egypt was as a factor in the internal political rivalries at Rome. In 64–63 Ptolemy sent an embassy to Pompey at Damascus; and at Rome Cicero prevented the passing of a land law that would have included the annexation of Egypt (in the interest of Crassus and the commercial class). At last in 59 Caesar was won over, and Ptolemy XII was recognized as king; Cyprus, however, was detached from Egypt and made into an independent province. In 58, a rising in Egypt caused the expulsion of Ptolemy, who was restored forcibly (for a large sum of money) by Aulus Gabinius, the governor of Syria, in 55. To achieve the restoration, Egypt was mulcted, its people were harassed, and many were murdered.

The succession of Cleopatra VII

On his death (51) Ptolemy Auletes left his kingdom to his children Ptolemy XIII (aged 10) and the famous Cleopatra VII (aged 17), as consorts. The ministers Pothinus and Achillas set the brother against his sister; and during her flight Pompey arrived in Egypt as a fugitive after his defeat at Pharsalus and was assassinated. Upon Caesar's arrival, Cleopatra returned to Alexandria and gained his support. In the ensuing civil war Caesar was hard pressed but eventually victorious, and Ptolemy XIII perished in 47. Cleopatra's sister Arsinoe later appeared in Caesar's triumph. Cleopatra, who was married to a second brother, Ptolemy XIV, went with him to Rome, where she remained until after Caesar's murder in 44. Some time later she replaced Ptolemy XIV as co-ruler with Ptolemy XV (Caesarion), allegedly her son by Caesar.

Within Egypt Cleopatra relied largely on the native element. She could speak Egyptian, an unusual accomplishment for a Ptolemy. It was also essential to have a strong friend at Rome. In 41 she was summoned to Tarsus by Mark Antony (Marcus Antonius), now in charge in the East; but the ensuing liaison was broken in 40 when Antony was reconciled to Octavian (later the emperor Augustus), whose sister Octavia he married.

Between 40 and 37 Roman armies in the East were fighting the Parthians, who had swept over Syria and Anatolia; in 38 the Romans drove them back to the Euphrates, and in 37 Herod, the son of the Idumaean king Antipater, expelled them from Judaea and established himself on the Judaean throne.

Antony resumed his liaison with Cleopatra, sending Octavia back to Italy; and in 37–36 at the so-called Donations of Alexandria, a last short-lived organization of the Hellenistic East took place. Asia and Bithynia remained provinces; but most of Asia Minor was assigned to the three augmented kingdoms of Pontus, Galatia, and Cappadocia (Cilicia Campestris was added to Syria, and the Lycian cities remained independent). Farther southward Nabataea and Judaea were client kingdoms, while Egypt was given southern Syria, Cyprus, and part of Cilicia Trachea. In 34 Cleopatra was designated queen of kings; Caesarion, king of kings; and the three children Cleopatra had borne to Antony were assigned satellite kingdoms—Armenia and lands beyond the Euphrates to Alexander Helios; the lands west of the Euphrates from the Hellespont to Phoenicia to Ptolemy Philadelphus; and Libya and Cyrenaica to Cleopatra Selene. These plans were destroyed in 31, however, when Octavian defeated the Antonian and Egyptian fleets at Actium. By 30 Octavian was in Alexandria and the Ptolemaic monarchy dissolved. Egypt was thenceforth a province of Rome, enjoying special relations with Octavian and his successors. The Hellenistic world had become part of the Roman Empire.

Hellenistic political, social, economic, and cultural institutions. *Hellenistic monarchy and royal administration.* Hellenistic monarchy united features taken from Macedonian kingship, in which the god-descended king claimed the personal allegiance of his peers, and from the absolute monarchies of Persia and Egypt. It built up a set of fairly homogeneous institutions, which varied slightly from state to state. The outward marks of Hellenistic monarchy common to all the kingdoms included dynastic rule, often with co-regency to ensure the succession; a divine pedigree; the use of the royal diadem, purple robes, and seal; the king's portrait on the coinage; and reckoning by years of the reigns. Dynastic intermarriage was normal and polygamy not unknown, but only in Egypt was brother–sister marriage usual. The wives of the kings frequently exercised great and independent power.

In the new realms (excluding Macedonia) the king was the state and his powers were absolute. His claim to this position rested on conquest of the land and his capacity to rule. But his position was reinforced by the adoption of ruler cult, an institution with mixed origins. To the native population of Egypt the pharaoh was always a god; but the official cults were something different and more Greek. In 324 Alexander had demanded worship from the Greek cities. The main step, however, was taken by Ptolemy II, who in 283 consecrated the dead Ptolemy I as the "saviour god," adding his wife to the cult on her death. In 270 he consecrated his own late wife Arsinoe II as the "brother-loving goddess," and he later set up a joint cult of himself and Arsinoe as the *theoi adelphoi* ("brother and sister gods"). His successors added themselves to this growing cult embracing all the Ptolemies; Alexander was set at the head, and Ptolemy IV incorporated Ptolemy I and his wife (the saviour gods) in the sequence. Royal cult was introduced into the Seleucid kingdom, probably by Antiochus I; and all the other kingdoms quickly followed suit. The Greek cities also instituted cults, separate from those of the states in which they were situated.

Mixed origins of the ruler cult

The precise meaning of such cults is unclear. God-kings were the object of vows but not of prayer. Yet the view that deification was designed to give the king a footing in the Greek cities is not supported by evidence. There was a strong political element, a means of uniting often very disparate areas and people in a common loyalty; the Greeks had always been familiar with gods who had mortal parents (like Apollo or Dionysus) or who had actually been mortal (like Asclepius or Heracles), and they were ready to equate achievement and service to mankind with godhead. The institution was later to be taken over by the Roman emperors.

The courts and administration of the various kingdoms shared certain features. The king had a council of friends (later elaborately subdivided into various grades) with advisory duties; these and the "kinsmen" were of Achaemenid origin. As the fount of law, the king possessed a secretariat to draft the decrees and letters in which law was formulated; and there was a secretary who wrote the official journal. From these functions a chancery developed with its own protocol and language sufficiently distinctive to influence the style of contemporary writers.

Other institutions common to most of the kingdoms were the guard (*agēma*), the corps of royal pages, and the bodyguards (*sōmatophylakes*), all inherited from Alexander. Normally there would be a finance minister (*dioekētēs*)—like Apollonius under Ptolemy II—and sometimes (perhaps only for an emergency) there might be a minister of affairs. The various districts within the boundaries of the state were usually governed by generals (*stratēgoi*) with military powers, responsible directly to the king (though the Antigonids used them rarely in Greece and not at all in Macedonia and Thessaly); in the Seleucid kingdom, down to the reign of Antiochus III, satraps were found, except in Asia Minor. There were admirals (navarchs) to control naval forces and, often, city governors (*epistatai*).

The army-assembly tradition

The traditional army assembly continued to be a factor in Macedonia, and some of its rights—the acclaiming of a new king, the appointing of a regent when the king was a minor, the approbation of the king's will and testament, and the judging of those charged with high treason—were sporadically exercised under the immediate successors of Alexander. Occasional political acts of the Seleucid and Ptolemaic armies in Antioch or Alexandria may be based on this Macedonian tradition, though often they seem to have acted more like a praetorian guard. Those armies ceased to contain any genuine Macedonian elements, except, of course, in Macedonia itself. In general, the administrative machinery was overcentralized, and, in consequence, far too much had to be done (or was left undone) by the king in person. Especially in Ptolemaic Egypt, the king faced an overwhelming volume of paperwork as petitions and references travelled up and decisions travelled down the chain of control.

The old and new Greek cities. Though really independent *poleis* were rare, the Greek city played a fundamental part in all the Hellenistic monarchies except perhaps Egypt. It was preeminently through the cities that Hellenism was maintained. King and city were interdependent, but their relationship was ill defined and rested ultimately on a basis of power. Although the machinery of an independent state was usually maintained through its council, assembly, and magistrates, royal domination was frequently assured by a king's party, by uniting several towns or villages into a single community to create virtually new cities out of old, and by the presence of royal governors and garrisons. The king's attitude toward the cities was based mainly on expediency, and the slogan of "liberation" was cynically exploited.

Ultimately only a city such as Rhodes, capable of military resistance, could be truly independent; for the rest, freedom from taxation was the nearest approach to liberty within their grasp. Many of the cities of Greece proper and the Aegean maintained a theoretical independence, which in practice was diminished by the presence of Macedonian garrisons at key points and, in some periods, by local tyrants in Macedonian pay. The development of leagues and other federal bodies compensated for this weakness (see above *The Greek leagues and the Antigonids in the 3rd century BC*) unless, as frequently, such bodies were themselves the instrument of a king (*e.g.*, the Islanders' League or the Hellenic Alliance of Antigonus Doson). Because Macedonian domination was always resented, the "liberation" of Greece by Flamininus in 197 was greeted with joy.

Under such restrictions men turned from politics to social and private life; clubs became more important than the state, and the main magistrates came to be those concerned with the food supply and the gymnasiums. The reduction in the number of small wars was a positive gain, an advantage that was increased by the kings' role in assigning foreign judges or making grants of asylum to shrines or even cities.

New Asian cities

The new cities in the kingdoms of Asia differed from the old city-states. Though often Macedonian military settlements in origin, they became centres of Hellenism, even if the racial basis of citizenship was less rigorously enforced. They were usually built on the gridiron pattern, and their population was divided in the old Greek way into tribes. As already mentioned, these cities played an important role in Seleucid Asia; in Egypt, they were rare and scarcely essential to the Ptolemaic pattern of government (apart from Alexandria, which was hardly a normal Greek city). The Attalids emulated the Ptolemies in maintaining a strict control of the cities within their dominions. But nowhere did the Greek cities have any say in policy; their function was to radiate Hellenism and to provide manpower reservoirs from which the kings could draw their administrators.

Hellenizing of non-Greek lands. Alexander's campaigns had initiated a period of Greek colonization comparable to the early migrations and to the movement of the 8th to 6th centuries; during the late 4th century thousands found new homes in Asia or Egypt, the more easily because the new political order facilitated movement. With them they took their political institutions, their language, and their way of life. Throughout the new monarchies Greek was the language of the army and the bureaucracy. In its Hellenistic form (the Koine) Greek became a world language, spoken from Spain to Afghanistan. Other Hellenizing institutions were Attic law, which was widely adopted in forms modified to suit local traditions, and especially Greek education. To have passed through the gymnasium—usually a private institution, though occasionally aided by royal endowment—was to have entered the gateway to Hellenism and material success. Out of the old military training (the *ephēbeia*) had also grown a new physical and intellectual regime designed to inculcate Greek values and produce Greek gentlemen and bureaucrats capable of serving in the administration.

For a century the Greeks, constantly reinforced from the homeland, remained exclusive in their culture. Then gradually upper class Asians became accepted and, when properly Hellenized, would live like Greeks and take Greek names. The values of the gymnasium were reinforced by the performances in the Greek theatre, which was to be found as far away as Babylon by the mid-2nd century. The guilds of travelling actors, on the fringe of upper class life—the *technitai* of Dionysus—found a new field of activity in the monarchies and travelled widely throughout the expanded Greek world. Greek, too, were the many private social clubs on which so much of men's interests now centred and that served as an avenue into society for foreigners arriving in a new city.

Class structure in Greece and eastward

For a few the new kingdoms brought great wealth. The higher officials in the army, the navy, or the court were well-off, indeed. In the cities of mainland Greece and the Aegean islands, wealthy landowners, shipowners, and owners of workshops employing slave labour were still important; but farther eastward the upper layers of society were filled rather by the professionals employed by the kings to run the state. Professionalism was strengthened in many fields, such as warfare, by the availability of technical treatises. As a result, it took the place of the amateurish organization of the city-state, as befitted an age that fostered the systematic recording of knowledge and experience. Both the upper class and those immediately beneath were Greek or Hellenized natives who made a substantial livelihood as settlers, merchants, craftsmen, and soldiers, or practiced the profession of architect, engineer, doctor, or teacher. These people formed the active middle class in the new Greek cities, and social and economic life in the new lands rested largely upon that class.

At the bottom were the labouring classes—usually made up of the native population, whose status varied from place to place; they could include freemen, serfs, and slaves. They were mainly non-Greek (though mixed populations were found, especially in Syria and Egypt), and their status depended on the tradition of the various lands. In general, however, they were poor and oppressed, though

there was no racial or nationalistic policy directed against them. Their hostility and, in Egypt, their frequent resort to flight were reactions to the hardship inherent in their situation. Slavery was not especially significant, except in Greece proper; there were domestic and temple slaves. Serfdom was far more widespread in both Seleucid Asia and Ptolemaic Egypt. In a society where all were the king's subjects, this lack of personal freedom was not, however, greatly important.

Economic life. Until the campaigns of Sulla in 87–86, the upper classes in Greece enjoyed considerable economic prosperity; not all areas, however, shared equally in this. Athens fell behind its earlier standards of wealth until it recovered in the late 2nd century; and Sparta was rent with social struggles. Northern Greece prospered, especially Aetolia and Ambracia and the port of Pagasae under Macedonian control; and evidence points to expanded trade and a high standard of living in the cities of Asia Minor, where riches were far greater than anything recorded from Greece proper. This prosperity is reflected in the amount of tribute raised by Rhodes from its mainland possessions, in the size of gifts or dowries or subscriptions to clubs in a variety of cities, and not least in the amount of plunder eventually seized by the Romans. Evidence from Corinth and Messenia shows mainland Greece also sharing in this prosperity, though to a lesser degree. A progressive fall in interest rates in the early 3rd century may be a partial indication that money was plentiful. But while their upper classes were rich, the cities themselves were often poor and living from hand to mouth. Most needs were met by raising loans (which were often tardily repaid), and public works frequently languished. A property tax was rare; cities preferred to raise funds by indirect taxes on imports and exports, pasturing fees, harbour dues, and the like.

Conditions were less good for the little man and the artisan. For the free worker available figures suggest hard times in the 3rd century, as money depreciated in value, and wages failed to keep abreast of rising prices for food and housing. The gap that thus grew between rich and poor created unhealthy tensions; and many cities tried to ease conditions for the indigent by distributing free grain. Although strikes were virtually unknown, revolutionary risings occurred in several cities. The call for debt cancellation and land redistribution was sometimes reinforced with demands for the liberation of slaves (to serve as troops in the revolution). Such a pattern is most clearly evident at Sparta under Agis and Cleomenes and later Nabis, as well as elsewhere, including Achaea around 150–146.

Conditions within the Hellenistic kingdoms

Within the Hellenistic kingdoms conditions were different. Those regimes were organized as largely closed economies with a high degree of governmental control. Especially the Ptolemies, themselves active in commerce, operated a strictly centralized economy outside the operation of the market; ultimately, against the heavy burden imposed by the bureaucracy, the only remedy open to the peasantry was flight. The other monarchies, to a lesser extent, employed a similar system. In Asia much of the wealth released by Alexander's campaigns together with other capital went to promote trade farther eastward until the initial impulse died down after about 250. In Egypt in the 2nd century, inflation was widespread and contributed to a growing inefficiency in the bureaucratic machinery. The centralized controls within the monarchies allowed agriculture to benefit from improved techniques, more intensive exploitation of the land (which was sometimes added to by draining, as in the Fayyûm in Egypt), and the introduction of new plants and crops together with a few new domesticated animals. Large estates, devoted to the enrichment of a minority, also afforded opportunities for experiment and more scientific farming. Though a few large slave-manned establishments were found in such cities as Pergamum and Alexandria, industry remained (as in the Greek cities) on a small scale; as in agriculture, the main advantages went to the government through heavy taxation or through the licensing of monopolies for the production of such commodities as salt, oil, papyrus, and grain. Trade was burdened with heavy customs dues, and foreign trade was largely in luxury goods and slaves. Production for the market was rare.

The declining economic life of the Hellenistic world suffered heavily from the ravaging that accompanied the Roman conquest. From 150 onward, the growth of piracy and slave dealing further disrupted economic life, and recovery had to await the Augustan peace and the incorporation of Egypt, Asia Minor, and Syria into the Roman Empire.

Hellenistic culture. The Hellenistic world was culturally one: extending from the western Mediterranean to Central Asia, it was linked by the use of the Koine form of Greek and by new facilities for travel. A lively atmosphere of curiosity prevailed, and a large public, even though not highly educated, turned eagerly to popularized knowledge and to a wide range of reading, often of indifferent merit. Ideas circulated, and among the cultural minority controversy stimulated research. There was cross-fertilization between the various philosophical schools; and in the arts masters created their own following, often associated with particular cities. The barriers of the *polis* had broken down and with them the old division between Greek and barbarian. In the ferment of such centres as Antioch, Pergamum, and Alexandria, the East made its contribution no less than did the old Greek cities such as Athens.

Literature and scholarship

A great role was played by royal patronage. The library and museum at Alexandria fostered literature and scholarship: the systematization of all kinds of knowledge, already pursued enthusiastically within the school of Aristotle, was there developed into a science. Modern texts still owe much, both for good and bad, to the Alexandrine philologists, whose prototype is perhaps the famous Didymus (*c.* 83–10 BC), author of 3,500 books. A vast literature grew up in all the main areas of the Hellenistic world; almost all of it is lost, but some of the New Comedy plays of Menander (*c.* 342–*c.* 292 BC) are among the works that have been recovered on papyri from the sands of Egypt. Writers showed a growing interest in personality, reflected in such works as Theophrastus' *Characters* (*c.* 300 BC) and a host of biographies, which culminated in the *Lives* of Plutarch in the 1st century AD. The end of political independence led to a diminution of political oratory but not rhetorical studies. The sensational and dramatic aspects of the historiography of the 3rd century came under heavy criticism in the next century by Polybius, the historian of Rome's rise to world power. New genres appeared, such as literary and cultural history and art history; chronology was scientifically studied, and Polemon of Ilium (2nd century BC) made use of inscriptions. The widening of frontiers brought with it a growing interest in geography, both descriptive and scientific; here the works of Eratosthenes (*c.* 276–*c.* 194 BC), Pytheas (*c.* 300 BC), and Polybius (*c.* 200–118 BC) were to lead on through Poseidonius (*c.* 135–*c.* 51 BC) to the work in Augustan times of Strabo. Vast works of compilation and encyclopaedias were balanced at the other end of the scale by romantic novels and scurrilous works of scandal.

In the visual arts there was a similar variety. A general restlessness encouraged change and the exploration of new concepts. Sculpture lost its formal composure and became tortured or realistically down-to-earth, portraying the individual rather than some idealized type. Romantic themes occur, and there is an interest in children. Famous schools grew up at Rhodes that were lively and technically very competent; of Chares' "Colossus," unfortunately, no clear account survives. The school of Pergamum, deriving largely from Scopas (4th century BC), is famous for such violent and emotional pieces as "The Dying Gaul" and above all the great frieze on the altar of Zeus depicting the battles of the gods and titans, the symbol of a world in struggle. Of painting only the faintest idea can be gained from such sources as the gravestones of Pagasae, the mummy cases of the Fayyūm, and a remarkable mosaic representing Alexander and Darius, probably at Gaugamela. Hellenistic painting was clearly an important art form, which was to be influential (along with sculpture) in fertilizing the arts of the East during imperial times.

Oriental traits in religion

Reactions to the new world after Alexander are reflected especially in the growth of Oriental traits in religion. The Olympians, associated closely with the old city-states, had ceased to mean anything to ordinary people; and attempts

to revive their cults, especially in the later 2nd century, proved ineffective. The Greeks, as they travelled and settled, sometimes brought their own gods along but often adopted those of the native people. Worship became more personal and linked with club life rather than the city. The combining of beliefs was widespread: there was felt to be one god, no matter how many names he bore.

The multitude of gods in Syria had a wide following, especially Atargatis (who was associated with Artemis); many were personalized cult objects, stones, or meteorites. Above all, men worshipped the great Anatolian nature goddess in her many guises. Of the older gods Dionysus was the most popular; and he was sometimes identified with Sarapis, the great creation of Alexandria, intended to link Greeks and Egyptians, who, however, rejected it. Isis, Sarapis, and Anubis formed a triad, worshipped outside and inside Egypt; Isis especially had perhaps the most widespread and far-reaching cult in the Hellenistic world and was later to influence Christianity through the figure of the Virgin Mary. But mystery religions that offered personal salvation flourished everywhere, as did magic, the shortcut to the achievement of one's immediate and often trivial desires.

Philosophy appealed to a different level of the population. Athens remained a great centre and gave hospitality not only to its own Academy (the school of Plato) but also to the Peripatetics (Aristotle's school) and later to Epicurus (341–270 BC) of Lampsacus and Zeno (*c.* 335–263 BC) of Citium, founders of the two most important Hellenistic schools—the Epicureans and the Stoics. After the work of Theophrastus and Strato, the Peripatos had lost its momentum and counted for little by 250; and the Academy under Arcesilaus took over much of the teaching of the Skeptics, who advocated the suspension of judgment, so that it became largely identified with the criticism of Stoicism.

Both Epicurus and Zeno were concerned with personal conduct rather than with abstract truth—with the individual rather than the nature of the outside world. They showed small interest in politics or indeed in the learning and science of their age. Both set out to equip men to deal with the blows dealt by fortune in an age of violent change—Epicurus by a doctrine of quietism and happiness, resting on virtue in a universe where the random movement of atoms was the ultimate causative force and the gods were remote and aloof; and Zeno by preaching the overriding force of universal law, the power of destiny, that man's greatest goal was to become the Stoic "wise man," and to live "according to nature." Of these, Stoicism was to live on, modified by Panaetius (*c.* 180–109 BC), into the Roman Empire.

Scientific development

At the other end of the spectrum, the age was one of scientific development and of progress in the arts that impinge on the sciences. Systematic work in botany, zoology, geography, ethnography, and hydrography was pursued in the Peripatos. The description of animals was furthered by the availability of zoos in such cities as Alexandria; and Theophrastus wrote a *History of Plants;* both zoology and botany were handicapped, however, by the lack of any scientific classification. Long controversy raged between supporters of the heliocentric and geocentric theories of the universe: the incorrect but more convincingly argued Earth-centred theory of Hipparchus (fl. between 146 and 127 BC) of Nicaea carried the day against the Sun-centred theory of Aristarchus (*c.* 270 BC) of Samos, and it was to persist until the Renaissance. Much headway was made toward creating a world map, despite the inability to solve the problem of detecting longitude. Eratosthenes, however, succeeded brilliantly in estimating the size of the Earth's circumference. Mathematics was developed by Eratosthenes and by Archimedes (*c.* 287–212 BC), who worked out the value of π and who also made a reputation (which he despised) as a practical mechanic, with such inventions as the screw for raising water. On the whole, although progress was made in siege engines, water clocks, eventually a water mill, a water organ, and the diopter (an improved form of water level), the Hellenistic world tended to agree with Archimedes in rating these activities below pure speculation. Some fields were thereby seriously handicapped—*e.g.,* physics and chemistry, which languished for want of optical glass. But progress was made in medicine, with increased knowledge of the nerves and arteries (though no one managed to detect and describe the circulation of the blood).

In one field, however, progress was linked up closely with the practical application of knowledge. The new cities gave wide scope for architects and engineers. The gridiron pattern originated by Hippodamus of Miletus (5th century BC) was everywhere reproduced (with appropriate concessions to the contours of the natural site); and the temples, marketplaces, and wide streets of Priene, Pergamum, or Miletus can still be reconstructed. Fortifications were devised to counter the increasingly effective artillery of the age; and many cities installed impressive water and sewerage systems, thus providing agreeable living conditions for large populations.

Commerce, travel, and exploration. Alexander's campaigns had widened the horizon of the Greek world, and his successors carried exploration further. Ptolemy II, in search of ivory and elephants, opened up the Red Sea along the Troglodyte coast and explored the Arabian coast (about 280); and the later Ptolemies pushed the boundaries of Egypt up the Nile to beyond Wadi Halfa and explored the African coast down to Somaliland. The Seleucids were active in the region of the Caspian, and farther eastward the Bactrian kings colonized Fergana and made contact with Chinese Turkistan. In the West, Pytheas of Massilia, probably a contemporary of Alexander, circumnavigated Britain and reached Norway or Jutland, but he met with widespread disbelief, among others from Polybius, who himself made a voyage in the Atlantic (146).

Trade by land and sea with India

Such journeys were undertaken partly in the interest of trade, and trade fostered knowledge of the outer regions. The Seleucids were interested in the route to India—by sea up the Persian Gulf to Seleucia, where it was joined by caravans coming overland from the Punjab through Taxila, Bactra, Hecatompylos, and Ecbatana; thence several ways led down to the coast, that by Antioch going forward along the old Royal Road through Tarsus to Ephesus. The Ptolemies encouraged the development of the route around southern Arabia, especially after Ptolemy V had lost Coele Syria; and toward the end of the 2nd century the collapse of the power of the Sabaeans in the Yemen opened up the way beyond. Eudoxus of Cyzicus was probably the first European to sail direct to India and back along the coast (*c.* 130 BC). Farther northward the Nabataeans at Petra built up a strong power by controlling the desert routes and sea trade in the Gulf of Aqaba.

Within the Helienistic world itself, trading was active. Grain from Egypt, Cyrenaica, and the Crimea (and later Numidia) came to the larger cities of Greece and Ionia, largely through depots at Rhodes and Delos. North Syrian and Ionian (and island) wine went everywhere. Textiles were produced in most areas that reared sheep, and silk from wild silkworms was a valued luxury; by the first century Chinese silk may have been reaching Egypt. There was an adequate supply of metals—iron in many parts, gold from new Nubian mines, silver from Asia Minor and Spain, quicksilver from Cappadocia. Many places had their specialties—Attic oil and honey, Bithynian cheese, salt fish from Byzantium, fruits and nuts from Pontus, and dried fruit from the Levant. Papyrus from Egypt vied with parchment from Pergamum. Glass, pitch, marble, bitumen, timber, granite, purple dye, and ivory were all well-attested articles of trade. But above all, it was the slave trade, centring in Delos from about 160 onward, that brought riches and misery as it grew along with Roman imperial expansion and the prevalence of piracy. (Delos, however, never recovered from the massacre carried out by Mithradates VI in 87.) Especially prized were the gems, spices, and frankincense from the East—from India and Arabia as well as from Palestine. For these Rhodes and Alexandria were important distribution centres.

Commerce contributed to the ferment of ideas as well as to the exchange of commodities and played a large part in the knitting of the Hellenistic world into a cultural whole. In general, this was a time when travel was not unusual. Among the many who travelled about the world were sol-

diers, especially mercenaries, professional men (teachers, philosophers, doctors, architects, and engineers) in search of patronage, actors, official envoys from kings and cities, some members of the royal courts and bureaucracies, and, not least, uprooted men (exiles, brigands, and pirates). But the majority, as always, stayed at home.

Macedonia and Greece under Roman rule (to c. AD 395). After 146 BC the decline of Greece continued. Social distress was accentuated by economic exploitation at the hands of the rich, both native and Italian. Athens was an exception, so long as it held the thriving port of Delos, but there were slave uprisings in Attica in 134–133 and 104–103, and in 115 Dyme was rent with class conflict. Sulla's campaign against Athens during the Mithradatic War brought widespread ruin, and during the 1st century BC the western coast of Asia Minor became the real centre of a Greek culture that continued to flourish despite political disasters. The civil wars between Caesar and Pompey and later those against Caesar's murderers and between Antony and Octavian bore hard on Greece: the battles of Pharsalus (48), Philippi (42), and Actium (31) were all fought on or around Greek soil. But under the early empire, Nicopolis Actia, Patras, and Corinth (refounded as a veteran colony) became flourishing centres of new life. Athens was in decline; but Sparta, under its tyrant Eurycles during the Roman civil wars, revived some semblance of its old Lycurgan regime, if mainly for the tourist trade. Delphi was impoverished, and the amphictyonic council counted for little until its revival under the emperor Hadrian (ruled AD 117–138). From 27 BC onward, southern Epirus, the Ionian islands, and the Cyclades were joined with Greece proper to form the senatorial province of Achaea under a proconsul of praetorian rank resident at Corinth and without a garrison.

Social distress and slave uprisings

On November 28, AD 67, Nero aroused wild enthusiasm by liberating Greece and freeing it from all tribute. The Peloponnese was renamed the Neronese, and Nero himself was worshipped as Zeus Eleutherios (Liberator). Nero's successors, the Flavian emperors, however, quickly reversed this gesture and restored taxation to the province. Prosperity revived to some extent under the emperors Trajan (ruled AD 98–117) and Hadrian, for although the upper class in Macedonia was somewhat mulcted to produce supplies for Trajan's wars in Dacia and Parthia, Greece generally benefitted from these conflicts and also from Hadrian's favour and enthusiasm for building. Athens received a whole new suburb, new buildings, and Panhellenic games were instituted; a new Athenian era was dated from the first of Hadrian's three visits in AD 124. The high-water mark of the revival of Athenian prosperity came under the Antonine emperors.

In the second half of the 2nd century AD the shadows lengthened. A plague brought back from the East by Lucius Verus' armies decimated Greece, and the barbarian Costoboci burnt Eleusis. Under Commodus and the Severi in the late 2nd and early 3rd centuries, tension between rich and poor grew greater and the institution of the *dekaprōtoi,* responsible for extracting taxes from an increasingly impoverished upper class, brought a final end to all pretense of free city institutions. In 251 the death of the emperor Decius opened the gates to barbarian invasion; in 267 the Heruli penetrated the Peloponnese and ravaged Corinth, Sparta, and Argos, until, after taking Athens, they were defeated by the historian Dexippus and a force of 2,000 men. Despite these disasters Athens saw some sort of revival coinciding with the teaching of Plotinus (*c.* 205–269/270) and the reign of Gallienus (ruled 260–268). Diocletian's reign (284–305) strengthened the central bureaucracy. Achaea and Macedonia, which now formed part of the diocese of the Moesiae, both decayed further. When the last Olympic festival was held in 393, few Greeks any longer took part. The whole peninsula had long been a backwater of small villages and large estates. (F.W.W.)

Ancient Italic peoples

Pre-Roman Italy was inhabited by peoples diverse in origin, language, traditions, stage of development, and territorial extension and was heavily influenced by neighbouring Greece, with its well-defined national characteristics, expansive vigour, and aesthetic and intellectual maturity. Italy attained a unified ethno-linguistic, political, and cultural physiognomy only after the Roman conquest; yet its most ancient peoples remain anchored in the names of the regions of Roman Italy—Latium, Campania, Apulia, Bruttium, Lucania, Samnium, Picenum, Umbria, Etruria, Venetia, Liguria.

THE ETRUSCANS

The Etruscans were called Tyrsenoi, or Tyrrhenoi, by the Greeks, Turskus by the Umbrians, and Tusci, or Etrusci, by the Latins. They called themselves Rasenna, according to Dionysius of Halicarnassus, and this is confirmed in the form *rasna* by Etruscan inscriptions. Their country was called Tyrrhenia in Greek and Tuscia, or Etruria, in Latin.

Location of Etruria

Etruria lay in central Italy between the valley of the Tiber, the valley of the Arno, and the Apennines, corresponding in present-day terms to northern Latium, a large part of Tuscany, and a tract in Umbria. The existence of the Etruscans in this territory is attested by epigraphic documents in Etruscan from the beginning of the 7th century BC onward. Other historical and epigraphic evidence attests to their settlements or their dominion in such other parts of Italy as Campania and Latium to the south and the plain of the Po to the north.

Origins. The problem of Etruscan origins has produced a discussion that has continued from classical antiquity to the present day. Its point of departure is the linguistic dissimilarity between the Etruscans and the other important historical peoples of ancient Italy who were Indo-European, such as the Latins and the Umbro-Sabellians. The problem remains formulated today in the terms in which it was set forth by Dionysius of Halicarnassus. Two opposite theses are commonly offered: (1) that the Etruscans came from the East by way of the sea (from Lydia, according to the tradition recorded by Herodotus, or in general from the Aegean and particularly from the island of Lemnos); and (2) that they were autochthonous (of local origin), or at any rate that they were an ancient Italian people (an early view, reviewed and critically formulated in modern times as a theory of their overland migration from the North or as a theory identifying them with the pre-Indo-European ethnic substratum of Italy).

Theory of overseas provenance from the East

The theory of the migration of the Etruscans from the East is the best known and most widely accepted by scholars. This provenance is based on the almost unanimous testimony of ancient tradition and on numerous other arguments—linguistic (onomastic, lexical, and structural affinities of Etruscan with the Asian and pre-Hellenic languages of the Aegean, especially with the inscription on an ancient funeral stele discovered at Lemnos), archaeological (diffusion in Etruria of the orientalizing civilization [influenced by the East]), and historico-religious (divination by means of reading animal viscera as in the Babylonian and Hittite world). It is generally believed and asserted that the Etruscans immigrated from the Aegeo-Asian area into Italy toward the end of the 8th or beginning of the 7th century BC, carrying with them a civilization of Eastern type. In such case the preexisting Villanovan iron culture would have to be attributed to Italic populations; but this conviction has begun to lose ground, for four basic reasons: (1) the fact that between the Villanovan culture and the orientalizing one there exists no clear break sufficient to attest the arrival en masse of a new population; (2) the disparate origins of the elements composing the orientalizing civilization of Etruria (from Egypt, Phoenicia, Mesopotamia, Urartu, Cyprus, and from Greece itself), which is more probably an indication of contributions by way of commerce and of cultural diffusion, as certainly happened in the corresponding cultural and artistic orientalizing phase in Greece; (3) historical and archaeological data, which demonstrate the presence of the Etruscans in the Italian area before the beginning of Greek colonization early in the 8th century BC; and (4) the chronology of the classical traditions, which places them in Italy in the 2nd millennium BC, at least four or five centuries before the migrations supposed by the modern theories. Some

contemporary scholars have indeed expressed the opinion that groups of Aegean navigators could have arrived in Italy at the end of the Mycenaean Period (*c.* 1200s BC), constituting the nucleus (or an essential component) of the future Etruscan nation. To them would have to be attributed the name of Tyrrhenoi, which may perhaps be identified with that of the Tursha, one of the "sea peoples" recorded in Egyptian monuments of the late 13th and early 12th centuries BC. (This line of thought envisions a type of "precolonization" similar to that of the Mycenaeans in southern Italy and in Sicily, or a phenomenon analogous to the maritime migrations and occupations by the Normans of the Middle Ages. In the latter case the Villanovan culture could be attributed to the Etruscan ethnos in course of formation and not to a pre-Etruscan ethnic stratum.)

Theory of provenance from the north

The theory of the provenance of the Etruscans from the north, already adumbrated by some 18th- and 19th-century scholars (simply because of the questionable comparison of the name Rasenna with that of the Raeti inhabiting the east-central Alps), has stressed as its basic argument the identification of the Etruscans with the Villanovans, held to be descended from a cremating people thought to have originated in central Europe. But discoveries in the 20th century show that the funeral rite of cremation was widely diffused by the end of the Bronze Age in the whole Italian area, including the South and Sicily (so-called proto-Villanovan), so that it seems difficult to attribute this phenomenon to a particular ethnic group. Entirely hypothetical also is the reconstruction of a Rhaeto-Tyrrhenian linguistic group of central European provenance with which to associate the origin of the Etruscans and of certain inhabitants of the northern Aegean, including the island of Lemnos.

Theory of the autochthony

Lastly, according to the autochthony theory, after the invasion by the Indo-Europeans (archaeologically identified with the cremators), the Etruscan nation survived and revived the most ancient populations of prehistoric Italy. A number of linguists attribute the Etruscan language to a supposed pre-Indo-European Mediterranean linguistic substratum common to Italy and the Aegeo-Asian area, thus explaining its affinities with the Anatolian and pre-Hellenic languages.

Each of these diverse theories contains positive and negative elements, forcing scholars to rethink in a broader way the question of the very terms in which the problem of Etruscan origins is stated and thus considering the historically more valid concept of a progressive formation of the Etruscan nation from the meeting of various external (Eastern and Northern) and local elements. This new perspective reduces the distance between the various traditional points of view to such an extent that it is possible to imagine a prehistoric connection of the Italian Tyrrhenic world both with the Greek and eastern Mediterranean area and with pre-Indo-European Italy. In any case, it seems certain that the formation of the Etruscan people should date from the transition from the Bronze Age to the Iron Age in about the 10th century BC, which is remarkably coincident with Etruscan beginnings according to some ancient traditions concerning the reckoning of the centuries.

Language and writing. Etruscan, the third great language of culture in Italy after Greek and Latin, does not survive in any literary works. Scholars know of the existence of an Etruscan religious literature, and evidence suggests that there may have been a body of historical literature too. But Etruscan had already ceased to be spoken in the time of imperial Rome, and its surviving writings interested only a few scholars, who made no translations or synopses. Only a few general notices and a few fragments exist today. Thus moderns will forever remain ignorant of the greatest part of the lexical and expressive inheritance of the Etruscan language.

The few known Etruscan documents are archaeological and almost exclusively of religious and funerary content. One of them, of exceptional interest, was found in Egypt, cut into narrow cloth strips wrapped around a mummy, and is now preserved in the National Museum at Zagreb, Yugoslavia. With about 1,300 words it is the longest existing Etruscan text; it contains a ritual of sacrifice, sufficient to give some idea of Etruscan religious literature. Found in Italy were an important religious text, inscribed on a tile at the site of ancient Capua, and an inscription on a boundary stone at Perugia, noteworthy for the singularity of its juridical content. Additionally there are a few thousand dedicatory inscriptions; grave inscriptions; and, more rarely, inscriptions of other kinds such as on coins, lead projectiles, dice, etc. The few Etruscan-Latin bilingual inscriptions, all funerary, have very slight importance with respect to improving knowledge of Etruscan. But inscribed gold plates found in Italy at the site of the ancient sanctuary of Pyrgi provide two texts, one in Etruscan and the other in Phoenician, of significant length (about 40 words) and of analogous content, that are the equivalent of a bilingual inscription and that offer substantial data for the elucidation of Etruscan by way of a known language—Phoenician; the find is also an important historical document that records the dedication to the Phoenician goddess Astarte of a "sacred place" in the Etruscan sanctuary of Pyrgi by Thefarie Velianas, king of Caere, early in the 5th century BC.

The modern concept of a "mystery" of the Etruscan language is fundamentally erroneous. There exists no problem of decipherment, as is often wrongly asserted. The Etruscan texts are perfectly legible. The alphabet derives from a Greek alphabet of Western type, disseminated in Italy by the colonists coming from the island of Euboea, probably during the 8th century BC, and adapted to Etruscan phonetics; from this same alphabet is derived, more or less independently, the Latin alphabet. (In its turn the Etruscan alphabet was diffused at the end of the archaic period [*c.* 500 BC] into northern Italy, becoming the model for the alphabets of the Veneti and of various Alpine populations, and concurred with the formation of the Umbrian and the Oscan alphabets in the peninsula.)

Etruscan alphabet

The real problem with the Etruscan texts lies in the cognition of the meaning of the words and of the grammatical forms. In this connection there exists a fundamental obstacle in that no other known language has close enough kinship to Etruscan to allow a reliable, comprehensive, and conclusive comparison. The apparent isolation of the Etruscan language had already been noted by the ancients; it is confirmed by repeated and vain attempts of modern science to assign it to one of the various linguistic groups or types of the Mediterranean and Eurasian world. But there are in fact connections with Indo-European languages, particularly with the Italic languages, and also with more or less known non-Indo-European languages of Western Asia and the Caucasus, the Aegean, Italy, and the Alpine zone, and on the other hand with the relics of the Mediterranean linguistic substrata revealed by place names. This means that Etruscan is anything but isolated and that its roots are intertwined with those of other recognizable linguistic formations within a geographical area extending from western Asia to east-central Europe and the central Mediterranean, and its latest formative developments may have taken place in more direct contact with the pre-Indo-European and Indo-European linguistic environment of Italy. But this also means that Etruscan, as scholars know it, cannot simply be classified as belonging to the Caucasian, the Anatolian, or Indo-European languages.

Methods of interpreting the Etruscan language

The traditional methods hitherto employed in interpretation of Etruscan were (1) the *etymological,* which is based upon the comparison of word roots and grammatical elements with those of other languages and which assumes the existence of a linguistic relationship such as to permit an explication of Etruscan from the outside (this method has produced negative results, given the error in the assumption); (2) the *combinatory,* a procedure of analysis and interpretation of the Etruscan texts rigorously limited to internal study of the texts themselves compared one with another and of the grammatical forms of the Etruscan words (this led to much progress in the knowledge of Etruscan, but its defects became evident in the hypothetical character of many of the conclusions due to the absence of external proofs or confirmations); and (3) the *bilingual,* based on the comparison of Etruscan ritual, votive, and funerary formulas with presumably analogous formulas from epigraphic or literary texts in languages

belonging to a closely connected geographic and historic environment, such as Greek, Latin, or Umbrian. But with the increase of reliable data, in part from more recent epigraphic discoveries (such as the gold plates at Pyrgi, mentioned above), the question of method appears to be of decreasing importance; that is, all possible procedures tend to be utilized, on the one hand maximally exploiting the various sources of external verification (ancient glossaries, bilinguals, parallel formulas in other languages, clues supplied by the nature of a monument and of its figuration, etc.) and on the other hand broad development of internal research on Etruscan, as through analyses and combinations of the contexts and of structural studies of the sentences and words.

The results already obtained from these investigations allow relatively easy and trustworthy translations of most of the shorter inscriptions and of many parts of the longer texts. It is important to note that Etruscan seems to differ from the structure of the Indo-European languages, such as Greek and Latin, and rather to approach in some aspects the structure of the agglutinative languages such as the Finno-Ugric.

Etruscan temperament and customs

Characteristics. Much has been said about the temperament of the Etruscans, as contrasted to that of the Greeks and of the other Italic peoples, particularly the Latins, stressing the aspects of irrationality, fantasy, mysticism, and discontinuity in Etruscan religious conceptions and artistic expression. A preeminence of women and of the feminine spirit is alluded to, in opposition to the masculine character of the expressions of Greek and Latin civilization.

With regard to Etruscan customs, there has been pointed out, among other things, the emphasis of the ancient writers on the Etruscan predilection for the flute, an instrument of Asian origin, on their orgiastic music, which accompanied not only their banquets but also the preparation of their food, the hunt, and even the flogging of slaves.

But in fact, the mentality and customs of the Etruscans can be explained as primitive survivals in a world started upon rapid intellectual, moral, and aesthetic advances by the powerful influence of Greek culture. In the Etruscan world there existed a contradiction between, on the one hand, prehistoric concepts and traditions and, on the other hand, the rapid and vigorous economic, technical, and political progress of the Tyrrhenian cities. While experiencing profound and determinant influences from Greek civilization, the Etruscan world remained extraneous to the rational order of classical thought.

The overall picture formed of the Etruscans in the 8th–6th centuries BC is that of a nation of navigators, traders, and industrial producers, driven to actions of immediate practicality; this also explains the fragmentary and transitory character of their political expansion. Their sudden and premature decline in the 5th and 4th centuries BC gives the impression of a cycle of power and of civilization that was never completed. In the last phases, beginning in the 4th century BC, the Etruscan nation seemed to withdraw into itself in its archaism and its provincialism, dominated above all by the elaboration of a complex religious thought and by the minuteness of its rites, which made it proverbially famous in the ancient world and an authority on religion in Rome itself.

Etruscan society

Organization. Etruscan society appears to have been founded on a system of families of patrilineal descent. The name of every free citizen consisted of two elements: (1) his individual name (*praenomen*), generally short and simple; and (2) his family or gentilitial name (*nomen*), common to all the members of the family group—generally formed, by suffixes, as a derivative of a simple personal name, the name of a god, a place, or the like. This onomastic formula was common to the Etruscans, the Romans, and to other peoples of ancient Italy. A nickname (*cognomen*), typical of Roman onomastics, appears in Etruscan inscriptions rarely and generally at a late date. Frequently added were the patronymic, or father's, name and the metronymic, or mother's, name.

The presence of the gentilitial name presupposed membership in a class that enjoyed civil rights. And in fact slaves and foreigners carried only a single personal name. Within the gentilitial system it is difficult to distinguish, on the basis of onomastic formulas, between families of the upper and lower classes, in the sense of the patricians and plebeians of the Roman state. Originally, or very early in the period of the massive economic flowering of the Etruscan cities, there seems to have existed a relatively equalitarian society. But within this social structure there probably began to emerge, or reemerge, particularly in the period of Etruscan decline, an aristocratic class that held all or most civil and religious power. Only at the end of the autonomous life of Etruscan cities (*c.* 2nd century BC) did there occur, probably under Roman influence, the massive influx of freedmen and slaves into the gentilitial system and hence into the enjoyment of civil rights. The Etruscan political order was based on the city-state, similar to the Greek *polis*, including an urban centre and a more or less extensive territory. The number and names of these independent state units in the most ancient times are not known; recent discoveries of early minor centres in the interior of Etruria indicate that they may have been more numerous than those known at the high point of historic times. These minor communities may have become progressively absorbed into the territories of the more powerful cities. From the 6th century BC onward, territorial organization and political and economic initiative were sharply concentrated in a limited number of large city-states, the names of which, starting from the south, are—along the coast—Caere, Tarquinii, Vulci, Roselle, Vetulonia, and Populonia and—inland—Veii, Volsinii (possibly modern Orvieto or Bolsena), Clusium (now Chiusi), Perusia (Perugia), Cortona, Arretium (Arezzo), Faesulae (Fiesole), and Volaterrae (Volterra). City-states

Notwithstanding their full political autonomy and cultural individuality, and quite aside from specific alliances, these cities were associated in a kind of religious and probably economic league, on occasion not devoid of political considerations. It epitomized the Etruscan "nation." According to the classical sources it was composed of 12 cities, or populi, which more or less corresponded to the major Etruscan cities listed above. The centre of the league was at a sanctuary near Volsinii, the Fanum Voltumnae, where the representatives of the various states assembled annually to participate in games and ceremonies and to elect a chief priest.

Government

The Etruscan cities were governed originally by monarchies presumably similar to those recorded for the most ancient period of Roman history, with sovereigns invested with complete military and religious power. Subsequently the power appears to have been entrusted to elective and temporary magistrates, as at Rome and in the Greek colonies. It is difficult to fix the dates and manner of the change from monarchies to republics. Epigraphic sources and historical tradition for the 5th century BC still testify to the existence of kings at Caere and at Veii, but it is not known whether these were traditional monarchies or personal powers similar to those of the Greek tyrants. The institutional change probably depended in part upon the transformations of the individual societies—especially in the great mercantile cities of the coast such as Caere, Tarquinii, and Vulci—under the influence of the Greek world, with which they were linked by very close ties of commerce and civilization (such centres as Sybaris in Magna Graecia and Miletus in Asia Minor; or Athens itself after Solon's reforms, which would mean during the 6th century BC). The evidence indicates that the republican system, as it appears in the latest period of the autonomous life of the Etruscan cities, was established in concomitance with the victory of powerful gentilitial and priestly oligarchies. This seems to exclude a process of evolution in a democratic direction, in spite of episodic revolutionary rumblings such as one in Volsinii in 265 BC. Only the direct influence of Rome and the general conditions in federated Italy, in the last days of the Etruscan city autonomies (in the 2nd and 1st centuries BC) and particularly in the cities of inland and northern Etruria, brought about a broad opening up to the lower classes.

Aspects of Etruscan civilization. The most conspicuous aspects of Etruscan spiritual and material civilization centre upon religion and the funerary realm. This does

not derive exclusively from the archaeological evidence of tombs and sanctuaries; tradition also attests to the predominance of the sacred sphere, especially in such aspects of Etruscan public and private life as politics, law, literary and artistic manifestations, and dress.

Etruscan cosmology

Etruscan religion. The peculiar tendencies of Etruscan religiosity are manifested in a sense of cancelling out the human vis-à-vis the often not readily knowable divine and in a continuous effort to interpret the will of supernatural beings by divination techniques and to honour it with scrupulous ritualism. The literary, epigraphic, and monumental sources provide a glimpse of a conception of cosmology whose image of the sky with its subdivisions is reflected in the consecrated areas and even in the viscera of animals. The celestial dome was divided into 16 compartments inhabited by the various divinities: major gods to the east, astral and terrestrial divine beings to the south, infernal and inauspicious beings to the west, and the most powerful and mysterious gods of destiny to the north. The deities manifested themselves by means of natural phenomena, principally by lightnings. They also revealed themselves in the microcosm of the liver of animals (typical is a small, bronze model of a liver found near Piacenza, bearing the incised names of divinities in its 16 outside divisions and in its internal divisions).

These conceptions are linked closely to the art of divination for which the Etruscans were especially famous in the ancient world. All public and private actions of any importance were undertaken only after having interrogated the gods; negative or threatening responses necessitated complex preventive or protective ceremonies. The most important form of divination was haruspicy, or hepatoscopy—the study of the details of the viscera, especially the livers, of sacrificial animals. Second in importance was the observation of lightning and of such other celestial phenomena as the flight of birds (also important in the religion of the Umbri and of the Romans). Finally, there was the interpretation of prodigies—extraordinary and marvellous events observed in the sky or on the earth. These practices, extensively adopted by the Romans, are explicitly attributed by the ancient authors to the religion of the Etruscans. The whole content of the rules of divination constituted a great corpus of doctrine that Latin authors referred to under the name of *disciplina Etrusca.*

The supernatural beings were extremely numerous, and many are unknown today. The cult had individual deities conceived in the same style as those of the Greek and Roman pantheon—*e.g.,* the supreme celestial divinity armed with lightning, Tin or Tinia (the equivalent of the Greek Zeus and the Roman Jupiter); the great mother goddess, Uni (Hera; Juno); perhaps Sethlans (Hephaestus; Vulcan); Turms (Hermes; Mercury); and others. There was no lack of Greek divinities directly introduced into Etruria, including Apollo, Artemis, Heracles (Hercle in Etruscan). The Etruscans populated the sky, the earth, the water, and the subterranean regions with an immense number of minor divinities and demigods, guardian spirits, nymphs, and demons. To the world beyond the tomb belong masculine (Mantus) and feminine (Vanth) divinities and those of uncertain sex—understood also as divinities of fate—and monstrous and terrifying demons.

Religious rites

The rites of the cult are substantially similar to those of the Greek and of the Roman religions, consisting of public and private sacrifices, with offerings of animals, liquids, and foods, accompanied by prayers and such other acts as observing the flight of birds and reading the viscera of sacrificial animals. Universally diffused was the custom of depositing in the temples more or less precious objects as offerings to the divinities, the recovery of which by archaeologists constitutes one of the most important sources of knowledge of the Etruscan religious conceptions, artistic taste, products, and material culture.

Etruscan funerary customs. The other principal source of evidence is, as already mentioned, the tombs. The idea of the survival of a dead person in the burial place involved the preservation and protection of his remains, the tendency to imitate the form of a house in the urns and sepulchres, and the custom of depositing alongside the deceased everything required for ultraterrestrial life, such as food and drink, clothing, ornaments, weapons, and furniture. Accordingly, the tombs reflect all aspects of the period from which they date. Most especially, the funerary pictures document Etruscan life, dress, and beliefs; the number of these documents, chiefly in Tarquinii has grown in the 20th century and by about 1970 amounted to about 100 painted tomb chambers. In the archaic period real-life subjects prevailed, such as scenes of banqueting, of music and dancing, of games and other funerary ceremonies, and of the chase. In the latest period (4th–2nd century BC) depictions connected with the world beyond the grave appeared. This signifies that new eschatological ideas had become diffused in the Etruscan world; the pessimistic tone of the later tomb pictures contrasts with the doctrine of the transformation of the dead into divinities, a doctrine reported in the literary tradition.

Cremation and burial

The rite of cremation of corpses, prevalent at the beginning in the Villanovan culture, remained characteristic of north-central Etruria throughout the duration of Etruscan civilization. The ashes of the deceased were placed in urns in the form of vessels, huts, and even of human figures. In the rest of Etruria the rite of interment prevailed; corpses were deposited in grave pits or upon beds in sepulchral chambers excavated in rock or constructed, later, within terra-cotta or stone sarcophagi, often carrying on the lid the image of the deceased.

Etruscan art and architecture. Unlike the other peoples of Oriental and classical antiquity, the Etruscans have left but sparse traces of a monumental architecture in stone; such building was limited almost exclusively to city walls and to some tomb structures. Temples and houses had only stone foundations and were built of unbaked bricks, wood, or other light materials. The wooden roofs were protected from the elements with sections of decorated terra-cotta; famous complexes of architectonic terra-cottas have been discovered at Veii, at Caere, at Falerii, and at Civitate near Siena. The form and the external and internal architectonic elements of the temples and houses are reconstructed through the evidence of their imitations in the tombs.

Arts of the Etruscans

The arts of the Etruscans originated in the traditions of the European and Mediterranean Iron Age cultures, with decorated objects and small plastic images conceived

By courtesy of the trustees of the British Museum



Painted frieze on terra-cotta slabs, 6th century BC, from Caere. In the British Museum.

with lively spontaneity or with geometric schemes. In the 7th century BC Oriental models established themselves, intermixing with the indigenous traditions. The influence of Greek art soon imposed itself in a preponderant and definitive way: first came the Daedalic and Peloponnesian styles; then the Ionic, particularly in the later part of the 6th century; and then the late-archaic trends, especially from Attica. From this there was an almost imperceptible transition to the protoclassical models of the first half of the 5th century. Archaic Etruria, in its greatest phase of power and wealth, is thus seen intimately linked to the image-making civilization of Greece. But the high level of quality and the technical refinement of some Etruscan works of art, especially in the period of Ionic influence (bronzes of Perugia; temple terra-cottas of Veii, including

From M. Pallottino, "Art of the Etruscans," Thames and Hudson, 1955, by Martin Hurlimann



Bronze relief from the caldron support known as the Loeb Tripod, *c.* 540 BC, from San Valentino, near Perugia; height 93 cm. In the Loeb Collection, Staatliche Antikensammlungen and Glyptothek, Munich.

a famous statue of Apollo; painted tombs of Tarquinii), indicates not a provincial or imitative art but an active participation by the Etruscan region in the archaic artistic culture. It is probable that Greek artists of some renown worked and created schools in the rich and powerful cities of Etruria. The creative and productive impulse continued with undiminished vigour through the first decades of the 5th century, so that Etruria welcomed the inspirations of the severe or protoclassic style.

Only later, when Greece and especially Athens had reached the peak of the inimitable creativity of classicism, did declining Etruria remain behind, bogged down in its traditions, impoverished in technique, disposed to welcome the inspiration of the classical models only slowly and only in their surface formulas. Beginning with the 4th century BC, the imitation of Greek art became an act of provincialism. Nevertheless, there were still late impulses of expressive originality, particularly in portrait art; and in this respect Etruscan art prepared to unite in a common Italic experience, which was the direct foundation of Roman art.

Etruscan expansion and foreign relations. During the archaic period the Etruscans constituted the most important political, economic, and cultural nucleus in the Italian area, excluding the Greek colonies of the south. Their dominion or control extended over most of the regions from the Alps to the Strait of Messina. More precisely, the ancient sources record a direct domination by the Etruscans in Campania and in the plain of the Po, with repeated and precise references to the Etruscan character of various cities in these territories (in Campania: Capua, Nola, Acerrae, Marcina, and others; in northern Italy: Felsina, Spina, Ravenna, and others; to these may be added evidence of a relative supremacy over Latium, with particular reference to Rome). The archaeological and epigraphic data confirm these references and permit scholars at least in part to state precisely the limits of the Etruscans' occupation and colonization and of their commercial and cultural influence.

Early expansion. The territorial expansion of the Etruscans is linked to their maritime power, of which the classical sources speak insistently. Etruscan presence overseas is recorded in such places as the Aeolian Islands and in general in the waters of Sicily, in Corsica, Sardinia, the Balearic Isles, and even on the coasts of Spain. A Greek source cited by Diodorus Siculus tells of a planned Etruscan naval expedition beyond the Pillars of Hercules (Straits of Gibraltar) to colonize a prosperous island in the Atlantic, perhaps Madeira or the Canaries; the expedition was not carried out because of the opposition of the Carthaginians. The memory of these exploits was embodied in the proverbial notoriety of the Etruscans as dreaded pirates.

Etruscan maritime power

Etruscan maritime activity in the Mediterranean coincided and competed with that of the Phoenicians and the Greeks. Its initial purpose was probably to extend and guarantee Etruscan commerce and its profits by controlling the seaways and creating markets in foreign lands. In these exchanges, metals mined in the ore-bearing zones of the western Mediterranean—particularly copper, iron, and silver—must certainly have been of primary importance. The Etruscans possessed rich mines (*e.g.*, near Vetulonia and on the island of Elba) and utilized them for their own exclusive advantage—for local needs or for selling and exporting the metals themselves—while jealously protecting these priceless sources of wealth against Phoenician and Greek cupidity. This explains the extraordinary economic development of some of the Etruscan coastal cities in the archaic period and the recurrent Greek push toward the ore-bearing districts of northern Tyrrhenia.

The wealth of the Etruscan cities is demonstrated by the enormous quantity and precious quality of their imports of Oriental and Greek goods between the early 7th and early 5th centuries BC; archaeological discoveries also reveal a widespread exportation of Etruscan products during this period to all the Mediterranean lands and especially to the west—principally vases of bucchero (a shiny black ceramic) and painted vases of Etrusco-Corinthian style. Thus the maximum Etruscan commercial and naval expansion must be placed between the later 7th and early 6th centuries BC, before the full predominance of the Greek navigators from Asiatic Ionia.

But Etruscan maritime activity probably began a good deal earlier; according to the historian Ephorus, Tyrrhenian pirates hindered the establishment of the first Greek colonies on the eastern coast of Sicily in about the mid-8th century BC, when Etruria was still in the Villanovan culture stage. Recent discoveries have revealed the Villanovan culture in Campania also, where the historical presence of the Etruscans is verified by literary sources and by Etruscan archaic inscriptions. These and other considerations indicate that the Etruscans may have preceded their conquest of Campania with a very ancient coastal colonization, in competition with the first Greek settlements of Pythecusae, Cumae, and Parthenope (now Naples). Campania would thus, by the dawn of historic times, have been about to become a zone of encounter between civilizations and, in certain respects, the key to the domination of Italy. In this framework, the very presence of the Villanovan culture in the plain of the Po could be evidence of an early Etruscan expansion beyond the Apennines, constituting the nucleus of the future Etruscan dominion of northern Italy.

Relations with the Greeks and Carthaginians. Sources indicate that the ancient maritime expansion of the Etruscans in the Mediterranean was an episodic phenomenon, prematurely arrested and then turned back and wiped out by more powerful activities of the Greeks and Carthaginians. Greek colonization in the central and western Mediterranean was strengthened beginning early in the 6th century by new commercial and migratory

Growth of Greek and Carthaginian power in the Mediterranean

currents from the coasts of Anatolia. In this same period Carthage was creating a vast commercial empire that absorbed all the area of Phoenician colonization in the west and precluded all further movement by the Etruscans in that direction.

The Etruscan relationship with the Greeks was twofold: it involved intensive commercial and cultural exchanges, from which derived profound and irreversible Hellenization of Etruria, and a basic rivalry for supremacy in the Tyrrhenian Sea and on the Italian peninsula itself, often intensifying into open conflicts. Greek–Etruscan relations may have resulted from diversities of interests of individual cities and from changing circumstances; but an antagonism of much vaster historical importance already was being established during the 6th century BC between the Greek colonies and Carthage, above all regarding the domination of Sicily. This led to a rapprochement between Etruscans and Carthaginians and to lasting ties of solidarity and alliance against the Greeks.

In this historic context were set various episodes, known fragmentarily from notices in the ancient sources. For example, there are records of repeated Etruscan attacks on the island of Lipara, colonized by the Rhodians but perhaps originally a lair of Tyrrhenian piratry against nearby Sicily and the Strait of Messina, as well as other naval actions of the Etruscans and the Carthaginians along the Sicilian coast. These were probably the last manifestations of an Etruscan presence in the southern Tyrrhenian Sea.

Reduction of Etruscan presence in southern Italy

But already toward the mid-6th century BC the Greeks had approached the coast of Etruria in founding the colony of Alalia in Corsica. A coalition fleet of Etruscan and Carthaginian ships battled the Phocaean fleet in the Sea of Sardinia (*c.* 540 BC). The consequence of this memorable battle was the Phocaean withdrawal from Corsica and the passing of the island into Etruscan control, while Carthage had a free hand for the conquest of Sardinia (effectuated in the late 6th and early 5th centuries BC). Notwithstanding this breathing spell, the Etruscan sphere of influence and of action within the northern Tyrrhenian Sea was thence forward reduced. In the late 6th century political control of Latium was lost with the fall of the monarchy of the Tarquins at Rome. If political significance can be attached to a gesture of homage by the king of Caere to the Phoenician goddess Astarte (according to the inscriptions on the Pyrgi gold plates), then even the Carthaginian allies themselves made their preponderance felt over Etruria. A further regression took place with the establishment of Syracusan hegemony in Sicily, which led to the defeat of the Carthaginians at Himera in 480 BC, and with a famous Etruscan naval loss (474 BC) in the waters of Cumae (inscribed on Etruscan bronze helmets dedicated by Hieron of Syracuse in the sanctuary of Olympia). The way was thus opened for repeated Syracusan naval incursions against the coast of Etruria, including the mining zone.

The loss of dominion on the high seas led to economic crises in the Etruscan coastal centres; the archaeological documentation shows a progressive diminution of foreign imports, ending with their near total disappearance and in the cultural and artistic decadence that characterized the 5th and 4th centuries BC. For this period some centres further inland, such as Veii, Orvieto, and Clusium, enjoyed a certain prosperity. But evidently, with the Tyrrhenian seaways closed to them, the Etruscans sought new outlets for their maritime commerce on the Adriatic Sea and new directions for land traffic toward the Alpine zones and toward central and northern Europe. There were sudden and impressive developments of Po valley and Adriatic Etruria, particularly the harbour city of Spina at the mouth of the Po and the so-called Certosa civilization at Felsina (now Bologna) and at Marzabotto in the valley of the Reno. A necropolis at Spina has yielded treasures of Attic painted ceramics of the 5th century BC.

Celtic incursions. But even this Etruscan vitality in the north was short-lived. The progressive infiltration and pressure of the Celts, who had penetrated and settled in the plain of the Po, eventually suffocated and overpowered the flourishing Etruscan urban communities, almost completely destroying their civilization by the mid-4th century BC and thus returning a large part of northern Italy to a protohistoric stage of culture. Meanwhile, the Gallic Senones firmly occupied the Picenum district on the Adriatic Sea, and Celtic incursions reached on the one hand Tyrrhenian Etruria and Rome (captured and burned about 390 BC) and on the other as far as Puglia. The Etruscan dominion in Campania was destroyed late in the 5th century by Italic Samnites descending from the mountains of Molise to the Tyrrhenian Sea.

Etruscan decline

The Roman conquest of Etruria. In the 4th century BC the picture of ancient Italy appeared profoundly transformed. As the protagonists of history, the eastern Italic people of Umbro-Sabellian stock diffused over most of the peninsula; the Syracusan empire and lastly the growing power of Rome had been substituted for the Etruscans (and for the Greek colonies of southern Italy). The Etruscan world had been reduced to a circumscribed, regional life, secluded in its traditional values; this situation determined its progressive passage into the political system created by Rome.

Probably beginning late in the 6th century BC and for at least a century under the Tarquins and other Etruscan monarchies or tyrannies, Rome experienced a deep Etruscan imprint and accepted extensive Etruscan elements into its population, reaching a level of economic and cultural splendour and of political power similar to that of the great Etruscan metropolises nearby, such as Caere and Veii. Etruscan influences related to such essential aspects of civilization as urban planning, architecture, art, religion, technology, and dress continued beyond Etruscan rule. Thus Rome, although remaining fundamentally Latin in language and in spirit, could yet be called by the ancients an Etruscan city.

Rome under Etruscan influence

The rivalry between Rome and the Etruscan cities, at least in the beginning, must have had the character of a contest between neighbouring countries participating in a common civilization, like the rivalry between Etruscan cities themselves. This was especially the case in the monarchical period, when Rome found itself under attack from Vulci and from Clusium. But already in the 5th century Rome gravitated toward the Latin world and became part of the Latin league, progressively assuming its leadership. Rome initiated a period of continuous border conflicts with closely neighbouring Veii lasting for almost a century and ending with the defeat and destruction of Veii and the inclusion of all its territory in that of Rome (396 BC). The Romans thus penetrated into the heart of southern Etruria, and for about another century the Etruscan states of this region faced alternating situations of amity and of war.

Roman policy was oriented toward control of Latium, of Campania, and then of regions farther south, with the aim of containing and eventually fragmenting the territorial power of the Samnites and their allies and of replacing them in hegemony over the peninsula; but the military conquest of Etruria probably did not constitute an interest or a necessity for Rome. Various Etruscan cities, especially in the north, such as Clusium, Perugia, and Arezzo, entered peacefully into alliance with the Romans. But when the Etruscan league, in conjunction with the activity of the Samnites and the Gauls (Celts), seemed to represent an effective and immediate danger to Rome (as in 311–310 and in 295 BC), the league's aspirations were quickly smashed. Rome also took Volsinii in 265 BC and deported its inhabitants, perhaps from Orvieto to the site of present-day Bolsena.

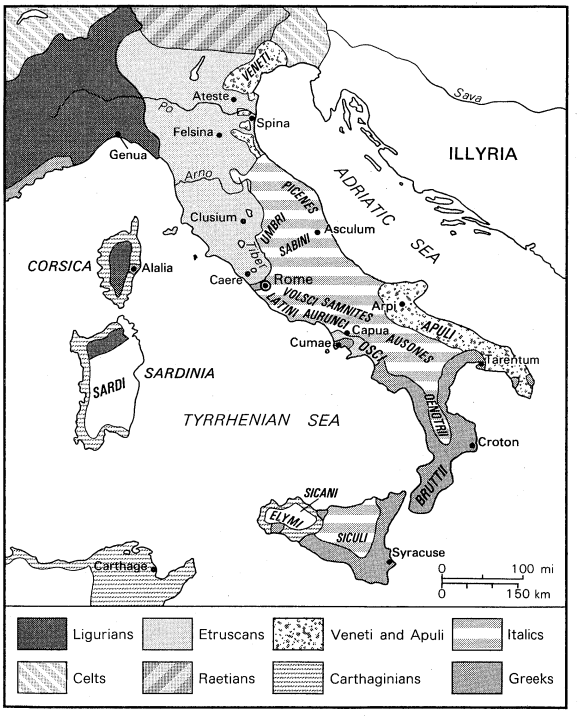
Alliances with Rome

By the mid-3rd century all Etruria appears to have been pacified and firmly subjected to Roman hegemony. Except for a few cases of annexation or special regimes, the Etruscan cities and their territories preserved a formal autonomy as independent states, with their own magistracy and other organs of government and their own religious and cultural traditions, until 90 BC, when Roman citizenship was extended to all the Italic peoples. This amounted to the complete political unification of the Roman-Italic state and to the disappearance of the last pretense of individuality and autonomy for the Etruscan city-states and for all the other allies of the peninsula. There followed a rapid process of romanization of Etruria and the replacement of its national language by Latin.

OTHER ITALIC PEOPLES

The Siculi

Local populations in areas colonized by Greece. The presence of the Siculi in Sicily and in the Italian peninsula is attested by the historical sources (Thucydides and Polybius). But the limits of their diffusion and their connections with other peoples of the peninsula (such as the Ligurians, the Itali, the Oenotrii, the Ausones) are more difficult to distinguish. A few small non-Greek inscriptions found in eastern Sicily and referable to the Siculi (the most noteworthy was found at Centuripe), coin legends, and Siculan words reported by classical writers demonstrate the Indo-European character of the Siculan language, which seems to show an affinity with Latin and also has connections with the Umbro-Sabellian dialects. The immigration of the Siculi from the Italian peninsula into Sicily goes back to a prehistoric but not extremely early epoch, considering that there are some echoes of it in tradition and that a continental archaeological influence suddenly appears at the end of the Bronze Age. The characteristic Siculan iron culture, evident in the necropolises of Pantalica near Syracuse and of Finocchito near Noto, flourished between the 9th and the 5th centuries BC and was progressively submerged in the superior civilization of the Greeks.



Distribution of peoples of ancient Italy *c.* 500 BC.

Evidence is quite scarce for the Siculi in the peninsula and for the other primitive indigenous populations of what is now Calabria and of Lucania (the Oenotrii, Ausones, Chones, Morgetes, and Itali) and Campania (the Ausones and Opici). Modern scholars have hypothesized that along the Tyrrhenian coastal arc there extended in earliest times

Western Italic peoples

a belt of paleo-Italic peoples (so-called western Italics or Proto-Latins) originally related to the Latins and distinct from the eastern Italic peoples inhabiting the Apennine and Adriatic regions. (The archaeological documentation consists of cemeteries with Iron Age graves, but there are also traces of cremation necropolises, especially in the province of Salerno and in Calabria.) Large fortified towns arose, particularly in the interior zones of Lucania. But the ethnic individuality of these ancient peoples was progressively obliterated between the 8th and the 4th centuries BC by the Greek penetration and by the expansion of the Etruscans and the eastern Italics (Samnites, Lucanians, Bruttii).

The Apulians. The inhabitants of the southeastern extremity of the Italian peninsula formed a definitely characterized group of populations that the ancients often called Iapyges (whence the geographical term Iapygia, in which "Apulia" may be recognized). The territory included the Salentini and the Messapii peoples in the Salentine Peninsula (ancient Calabia) and the Peucetii and the Dauni farther north. (Sometimes the designations Iapyges and Messapii are used with identical reference.) Ancient tradition insists upon an overseas origin for these tribes, held to be Cretan or Illyrian. The Iapygian, or more commonly, Messapian (Messapic) language is known from a considerable series of public funerary, votive, monetary, and other inscriptions written in the Greek alphabet and found in the Apulian area, especially in the Salentine Peninsula, from words reported by the ancient writers, and from toponomastic (local place name) data. Messapian is without doubt an Indo-European language, distinct from Latin and from the Umbro-Sabellian dialects, with Balkan and central European analogies. This confirms the overseas provenance of the Iapyges from the Balkans, the more so because there existed in Illyria a tribe called the Iapodes and because a people known as the Iapuzkus lived farther north, on the Adriatic coast of Italy. Rather than a true immigration, however, there was a gradual prehistoric penetration of trans-Adriatic elements. The expansion of the Iapyges must have brought them to Lucania and even to what is now Calabria, as would be deduced from traditional and archaeological indications.

Apulian civilization

The Apulians' civilization, which was considerably influenced by that of the nearby Greek colonies, developed from the 9th to the 3rd centuries BC. In the most ancient period there were pit graves, sometimes in large stone tumuli. In the Siponto area, near what is now Manfredonia, the graves were accompanied by anthropomorphic stelae with geometric bas-reliefs. Geometrically painted ceramics in linear motifs persisted to the threshold of the Hellenistic Age. Later graves took the form of large trunks and of catacombs with paintings on the sides. Burial was the disposition exclusively used.

Beginning in archaic times, large cities developed, linked to each other by bonds of confederation. These included Herdonea (now Ordona), Canusium (Canosa), Rubi (Ruvo), Gnathia, Brundisium (Brindisi), Uria (Oria), Lupiae (Lecce), Rudiae, and Manduria. They preserved their independence, tenaciously defended against the Greeks, until the age of the Roman conquest.

The Latins. The Latin nation had a relatively limited territory, south of the Tiber, which was reduced, in historic times, by the invasion of the Volsci to the region between the Alban hills and the Aurunci mountains (the so-called Latium Novum). The principal Latin centres included Alba Longa; Tusculum; Lavinium; Ardea; Tibur (now Tivoli); Praeneste (Palestrina); and the early Volscianized cities of Velitrae (Velletri), Signia (Segni), Cora (Cori), Satricum, Antium (Anzio), and Anxur (Terracina). The importance of the Latins is essentially linked with the fortunes of Rome, the forward bulwark of Latinity in the direction of the Etruscan realm. Intermixtures with the legends of the origin of Rome make the ethnographic traditions of Latium very diverse and complex. The linguistic evidence, which begins with inscriptions of the 7th to 6th century BC, indicates an individuality of the Latin world distinct from the neighbouring Etruscan and eastern Italic peoples.

The Latins had a federal organization, centred at the sanctuary of Jupiter on Albanus Mons. Their religious heritage survived in the beliefs and cults of the Roman world. The most ancient Latin culture (9th–8th century BC) was characterized by cremation as the funeral rite, in common with that of Etruscan and northern Italian territory, and by an iron culture showing affinities with the proto-Villanovan culture and with the cultures of Tyrrhenian southern Italy. Etruscan political control of Latium (probably 7th–6th century) coincided with an evident Etruscan cultural and artistic influence; while from the south and from the sea, elements of Greek civilization penetrated, beginning with the alphabet.

Latin culture

North of Latium lived tribes ethnically akin to the Latins, with principal centres at Capena, Narce, and Falerii (whence the name Faliscans). Their political and cultural history merges with that of the Etruscans. The Faliscan dialect, known from inscriptions, was originally Latin but

was contaminated and modified by eastern Italic and Etruscan elements.

The eastern Italics. A great part of the central and southern Italian peninsula was occupied in protohistoric and historic times by populations forming a vast ethnic and linguistic unit—the eastern Italics or Umbro-Sabellians. To the south, in the mountains of the Abruzzo, lived the Samnites, who later spread into Campania, Lucania, and what is now Calabria. In the centre were the Vestini, Paeligni, Marrucini, Marsi, Aequi, Volsci, and Sabini. Further north lived the Umbri. The origin and relationship of all these peoples is unclear. Ancient ethnographic traditions bearing on central Italy link the Samnites to the Sabines and the Sabines to the Umbri, locating their primitive centre of dispersion in the Rieti basin and in the area of Amiternum. Their diffusion was attributed to the mass emigration of an entire generation in search of a new homeland (so-called sacred spring).

Origin of eastern Italics

The linguistic data prove the unity of the group of eastern-Italic idioms, belonging to the Indo-European stock but differing from the Latin. Within this group may be distinguished a southern variant (Sabellic or Oscan) known from an abundant harvest of epigraphic documents from Samnium, Campania, and southern Italy. This variant is to be attributed to the Samnites and to a large part of the minor stocks of central Italy, including the Sabines, known from isolated inscriptions in central Italy. On the other hand a northern (Umbrian) variant is represented by inscriptions of Umbria—principally bronze tablets from Gubbio, inscribed between the 4th and 1st centuries BC by a brotherhood of Umbrian priests—and by a bronze tablet from Velletri. The eastern Italic words reported by the classical writers, as well as toponomastics, confirm these conclusions.

Notwithstanding the original unity of stock and of language, these populations had diverse histories and cultures. The Samnites from Molise (the Caraceni, the Pentri, and the Frentani) in the 5th and 4th centuries BC occupied Campania—where they vanquished Etruscans and Greeks and assumed from the local tribes the name Opici, or Osci—as well as Lucania (with the Hirpini or Lucani), reaching what is now Calabria—where they took the name Bruttii—and finally Sicily. Defeated by Rome in the Samnite wars (4th and early 3rd centuries BC), the Samnites tried for the last time, in the period of the Social War (90–83 BC), to counterpose to the Romans an Italic nationality of their own. A considerable difference existed between the culture of the mountain Samnites—organized in confederate tribes centred on fortified villages and in the 5th and 4th centuries still retaining aspects of the "iron culture"—and the high civilization of the Campani and Lucani established in the ancient cities of Capua, Nola, Nocera, and Paestum and dominated by Greek and Etruscan influence.

The Samnites

Some central tribes—the Marrucini, Vestini, Paeligni and Marsi—appear to be linked historically, politically, and culturally to the Samnites. The case is different with the Sabines, the Aequi, and the Volsci, whose period of expansion (6th and 5th centuries) is closely connected with earliest Rome and who had early contact with the Etrusco-Latin civilization.

The diffusion of the Umbri toward the north and beyond the Apennines lent credence to the ancient traditions relating to the great size of their territory. The traditions, however, are more probably based on the fact that the name Umbri is derived from that of a most ancient population, probably not Indo-European and certainly not Italic, living in the Apennine region before the diffusion of the eastern Italics. The history of the Umbrian cities—Iguvium (now Gubbio), Hispellum (Spello), Spoletium (Spoleto), Tuder (Todi), and others—is known only beginning with the period of the struggle of the Etruscans and Gauls against Rome. Umbrian civilization is revealed by the Gubbio Tablets, a document unique in its kind. The Umbrian artistic and material culture derived in large part from that of Etruria.

The Umbri

The populations of the Picenum. In historic times expansion of the eastern Italic peoples placed them firmly along the Adriatic coastal tract corresponding to what is now the Marches region. Epigraphic and archaeological data give evidence also of the presence in the Picenum (an ancient region between the Apennines and the Adriatic) of the trans-Adriatic Liburni and the pre-Indo-European Asili. It is possible that there were established here elements that had come by sea from Illyria (the Gubbio tablets mention the Iapuzkus, whose name recalls that of the Illyrian Iapodes and of the Apulian Iapyges). But inscriptions in the southern Piceno seem to exhibit a close kinship with the Umbro-Sabellic dialects.

A material civilization flowered between the 8th and 5th centuries in the northern Abruzzo and in the Marches. This civilization is represented by the rich funerary equipment of burial tombs, whose type and decoration present affinities with the iron culture of Tyrrhenian and northern Italy and with that of the Balkans, and show Greek influence. Cremation tombs of Villanovan type have been found at Fermo. Also noteworthy is the presence of stone funerary sculpture. North of Ancona is a cultural variant, particularly in the necropolis of Novilara near Pesaro, where inscriptions are in a dialect other than that of the southern Picenum and difficult to classify.

Picenum civilization

It may be held that the middle-Adriatic iron cultures expressed an early archaic mixture of eastern Italic and trans-Adriatic peoples, influenced by the Etruscans and the Greeks. Contributing to their decline were the Gauls and Syracusans, who established themselves in this area in the 4th century BC. In the 3rd century the Picenum was already totally conquered by the Romans.

The Veneti. Ancient tradition held the Veneti to be an Illyrian people who, coming from the east, took possession of the region named for them (Venetia). To them are linked the Histri, the Carni, and various Alpine tribes. (The name of the Veneti, or its root, is widely diffused in the ethnic anomastics of central Europe and even of Asia.) The Venetic language is known from funerary and votive inscriptions, from words cited by the classic writers, and from onomastic and toponomastic data. It is an Indo-European language of archaic type, with similarities to the Latin and the Germanic.

Venetic origins

The principal centres of the Veneti, located at the western margin of the territory, were Padua and Este. Their culture developed from the 9th century to the period of romanization, with relationships with the Golasecca, Villanovan, and Etruscan cultures and with the transalpine Hallstatt culture. Maximum development occurred in the 6th–4th century BC; particularly noteworthy is the production of figured bronze *situlae* (conical vessels). In the final period Gallic (Celtic) influences are found. Phenomena parallel to those of Este appear in Istria. The Veneti were horse breeders and peaceful traders and navigators; protected by the waters of the lower Po and the lower Adige, they preserved their independence against Etruscan expansion and Celtic invasion and in the 3rd century BC entered into peaceful alliance with Rome.

The Ligurians. For the ancients, the name Ligures designated the peoples of northwest Italy, including northern Tuscany, Liguria, Piedmont, and part of what is now Lombardy. Historical tradition also placed them in central Italy, while the classical writers and toponomastic affinities give them a broader diffusion beyond the Alps. The Ligurians included also the peoples of Corsica. The more ancient Greeks gave all the peoples of the western world the common designation of Ligyes (*i.e.*, Ligurians).

Linguistic data—furnished by toponomastics, lexical survivals, and by Ligurian words cited by classical writers—betray the presence of a pre-Indo-European Mediterranean stratum akin to that of western Europe. Inscriptions found in upper Lombardy and in the Ticino exhibit Indo-European characteristics and in particular Celtic influences. Thus the Liguri seem to belong to an environment formed in northern Italy after the Celtic invasion and called Celto-Ligures.

The Etruscan expansion in the plain of the Po and the invasion of the Gauls confined the Ligurians between the Alps and the Apennines, where they offered such resistance to Roman penetration that they gained a reputation with the ancients for primitive fierceness. Among the more considerable Ligurian monuments are rock engravings

Ligurian culture

and anthropomorphic sculptures analogous to those of southern France, found in Lunigiana and Corsica. Some of these artistic manifestations are repeated in territories farther east. But it remains doubtful whether the similar cultural imprint indicates an original identity of stock. Ligurian and Celto-Ligurian tombs of the Lombard lakes region, often holding cremations, reveal a special iron culture called the culture of Golasecca; while Ligurian sepulchres of the Italian Riviera and of Provence, also holding cremations, exhibit Etruscan and Celtic influences.

Populations of central northern Italy and of the Alps. The ethnography of the Po and Alpine regions is complex and obscure because of the early spread of Etruscan culture and colonization. The ancients record two major ethnic groups (aside from the Etruscans and the Veneti): the Euganei, inhabiting the plain and the Alpine foothills, and the Raeti, in the valleys of the Trentino and the Alto Adige. Minor peoples in the region belonged to one or the other of these stocks or to Ligurian stocks; with regard to many of these peoples the sources speak of an Illyrian or an Etruscan origin.

Late inscriptions discovered in the Adige River valley and on the plain have a dialect showing some affinities to Etruscan. Some scholars see in this a blending of local and Etruscan elements, while others speak of an indigenous pre-Indo-European language with Indo-European influences. Primitive toponomastics confirm the existence of a linguistic stratum that could be defined Raetian or Raeto-Euganean but distinguish it sharply from the Venetic and probably also from the Ligurian. Other inscriptions from the Val Camonica and the Garda district attest to a more noticeable Indo-European dialect, due perhaps to Celtic and Latin influences. To the west are the so-called Lepontian inscriptions.

Original central- and sub-Alpine popula- ions

Thus in the central-Alpine and sub-Alpine area, there were original populations, different from the Veneti and the Etruscans, whose kinship with the Ligurians remains uncertain. The distinction between Euganean and Raetic tribes can be based only upon an approximate geographical criterion. To this original ethnic stratum may have belonged the most ancient inhabitants of the region, before the immigration of the Illyrian Veneti and the Etruscan conquest; certain cremation sepulchres of the Verona and Mantua regions may be attributed to them. Perhaps the existence of a Venetian goddess Reitia, recorded by Strabo and mentioned in inscriptions from Este, is some proof of a Raeto-Euganean cultural persistence in the territory occupied by the Veneti. (Ed.)

Ancient Rome

From humble beginnings in the 8th century BC as a minor people settled on a few hills overlooking the Tiber River, the Romans progressively conquered the Italian peninsula, extended their dominion over the entire Mediterranean basin, and expanded their empire into continental Europe toward the Atlantic. Their conquests allowed them not only to dominate but also to civilize a large number of peoples, on whom they made a lasting impression. Modern historians can still distinguish those regions that were deeply penetrated by the Latin language and Roman law from those that, because of their remoteness, were untouched by it. In attempting to explain the extraordinary and lasting success of Roman culture, it is necessary to seek its sources not only in the vicissitudes of history but also in the psychology of the Roman people. Although the Romans were not gifted in the areas of artistic creativity, scientific research, or philosophic thought, they did possess a steadfast attachment to their native land and a profound sense of political and administrative organization. From a nation of peasants and soldiers, Rome was able to develop an urban civilization, on the pattern of the Greek *polis*. In the domain of art and thought, the Romans were the heirs of the Greeks, and their political genius, common sense, and exemplary piety—which gave them the constant conviction of being sustained and guided by the gods—made it possible for them to create an empire in which the conquered, instead of being kept at a distance, participated in the life and prosperity of the conqueror. No barriers were raised between the nations or races; the emperors themselves often came from outlying regions of the Roman world. The Romans' tenacity, civic feeling, and fundamental liberalism made it possible to establish a lasting peace among men and to confer the benefits of culture and material prosperity on a large number of peoples. Certainly, faults were not absent from a world that could not survive without slavery and that occupied its leisure hours with the unwholesome pleasures of the amphitheatre. The overall balance remains positive, however, and Rome appears to modern eyes as the great teacher of the Western world.

ROME FROM ITS ORIGINS TO 264 BC

Early Rome to the 6th century BC. *Historiographical problems.* For the history of the later republic and the empire there is an abundance of documentary sources, epigraphic texts, and archaeological data that provide the historian with a sound basis for research. But for the founding of Rome and the early years of the republic, there is little evidence. The archaeological remains for this period are rare, and inscriptions, which do not begin to increase in number until the 1st century BC, are scarcer still. There are, however, numerous Greek and Latin writings dating from the end of the republic and the beginning of the empire that recall the birth and early years of the *urbs*. These later histories have for centuries enriched the literary traditions of the Western world, but their value as history remains problematical. Until recently, critical opinion has been suspicious of the highly colourful accounts of the Roman annalists and has seen in them only fantasy and lies intended to appease the vanity of great families and the patriotism of a proud people. The attitude of many contemporary historians, however, is considerably different, for archaeology has brought indisputable corroboration to tradition. The tale of the founding of Rome by Romulus in the middle of the 8th century BC, for example, has to a certain extent been confirmed by the discovery after World War II of a village of huts on the Palatine effectively dating from about 750 BC. The power and prosperity accorded to the Rome of the Tarquins (Etruscan kings) by the annals are confirmed by discoveries resulting from excavations of ancient Rome. Archaeological and epigraphic discoveries, in fact, continue to demonstrate that the legendary guise of the traditional material actually masks a real foundation of authentic events.

Greco-Latin historiography

The myths of origin and early monarchy. The legend pertaining to the origins of Rome is twofold and distinguishes two events: the colonization of Latium, the territory surrounding Rome, in the 12th century BC by the Trojan Aeneas and the actual founding of the city of Rome 500 years later by the Latin Romulus. The arrival on the Latin plain of Aeneas and his troop of Trojans, who had escaped the destruction of their country, has until the latter half of the 20th century been viewed as pure legend. Contemporary archaeology, however, has proved that in that remote epoch Italy was visited by Late Bronze Age navigators whose trading activities have been attested as far as southern Tuscany. The legend of Aeneas perhaps expresses the memory of these ancient contacts with the Orient. In any case, the legend of Aeneas can no longer be viewed as the creation of Roman writers of the last centuries of the republic. The figure of Aeneas was familiar to Etruscans from the 6th century BC, and the legend could have made its way to Rome during the period of Etruscan hegemony.

Aeneas and Romulus

The second aspect of the legend concerns the founding of the town by its eponym Romulus, whose life is depicted in legend as a succession of marvellous and violent episodes. Having miraculously escaped death with his twin brother Remus, he and his brother asked for signs from the gods to designate which of them was to found a new city. Romulus was designated by the divine omens and killed his twin in a murderous quarrel. He then became the first of seven legendary kings who, according to tradition, ruled in Rome from *c.* 754 BC to the advent of the republic in 509 BC.

This account was elaborated gradually and had received its definitive form by the 6th century BC. It rests on a

The early kings

fact now attested by archaeology: the establishment on the Palatine (one of the seven hills of Rome) in the 8th century BC of a primitive village of huts where shepherds of Latin origin had gathered. Much of the legend surrounding the early history of Rome is derived from the account of Livy (59 or 64 BC–AD 17) writing hundreds of years after the events he describes. Livy's first book of Roman history recounts in great detail the activity and achievements of the kings. Mythical, legendary, and strictly historical elements are blended into the text of this narrative, which was based on the annal material. According to Livy, Romulus was a military chief who possessed wide political and military powers. He set out to organize the Roman state and, in order to provide wives for his comrades, planned the rape of the Sabines, a neighbouring people. The war with the Sabines ended in the union of the two peoples, and, after the mysterious death of Romulus in 717, a Sabine, Numa Pompilius, reigned until 673. He was a pious sovereign who dedicated himself to giving Rome, founded by force and arms, its laws and morality. It was he who first gave form to the Roman religion. He was succeeded by a warrior king, Tullus Hostilius (traditional reign 672–641), who framed the military code. The famous episode of the Horatii and Curiatii, the champions of Rome and Alba Longa, respectively, occurred during a war Tullus Hostilius conducted against Alba Longa. The line of pre-Etruscan kings ended with Ancus Martius, a Sabine, who was the grandson of Numa Pompilius. He supposedly expanded Rome and founded Ostia. Beneath the surface of the traditional account rests the reality of the progressive growth of a town that initially was only a federation of villages perched on the Palatine and neighbouring hills. Post-World War II archaeology has made it possible to trace from early times the course of the growth of Rome, which became a true city only with the arrival of the Etruscans. The Latin population settled on the hills overlooking the Tiber presumably brought with it its language and its scrupulous piety, expressed in a whole series of ancient rituals that were faithfully preserved until classical times.

Archaeological accounts

The Etruscan hegemony over Rome. According to tradition, the end of the 7th century and the 6th century BC were taken up by the reigns of the Etruscan sovereigns Tarquinius Priscus and Tarquinius Superbus, between whom was interposed a man of obscure origin, Servius Tullius. The annals give dramatic details of this period, during which Rome obeyed rulers of foreign origin who allegedly made the very name of royalty odious to the Romans. As legend has it, Tarquinius Priscus was a native of Tarquinii, where his father Demaratus, a Corinthian Greek, had taken refuge. This first of the Etruscan kings supposedly ruled from 616 to 579 BC. He was a good administrator and a great builder and gave Rome new lustre. Assassinated by the sons of Ancus Martius, who considered him a usurper, he was succeeded by Servius Tullius, whose name suggests one of humble origin. Many historians believe that Servius was an interpolation by later annalists, who refused to ascribe the achievements of this period to the Tarquins. Servius was an enlightened king who was supposed to have established the organization of the city by census results and the distribution of citizens into classes, in turn divided for military purposes into centuries (groups of 100). Servius is alleged to have divided the town into four territorial tribes and encircled it with a strong, unbroken wall. Assassinated by the sons of Tarquinius Priscus, Servius was succeeded by one of them, Tarquinius Superbus, whom legend depicts as violent and unjust. He successfully waged war against neighbouring peoples and continued his father's construction work. The Cloaca Maxima (a remarkable drainage system) and the completion of the temple of Jupiter Capitolinus are attributed to him. According to legend, the excesses of the Tarquins provoked an insurrection among the Roman citizens. Lucretia, an honorable Roman matron, supposedly committed suicide after being violated by one of the King's sons. Lucius Junius Brutus, a freedom-loving Roman, roused the outraged populace and condemned the tyrant and his family to exile. Thus was the Roman Republic born and power transferred to its legitimate custodian, the people themselves.

The reign of Servius Tullius

This fictionalized and patriotic account contains some elements of truth. Rome came under the ascendancy of its powerful neighbour Etruria in the 7th century BC. Somewhat later than tradition would have it, near 550, it was ruled directly by the Etruscans, who thus obtained an invaluable passageway to Campania, which they controlled. As a result of the Etruscan seizure of power, Rome became a great and prosperous city with a preponderant influence over the rest of Latium. The Etruscan leaders encircled it with a continuous wall and adorned it with stone monuments, the most imposing of which was the Capitoline temple, which was dedicated to Jupiter, Juno, and Minerva. The gods were Roman, but their grouping in a triad reflects the Etruscan predilection for grouping things in threes. Rome was thus genuinely affected by the Etruscan religious and cultural influence. The Etruscans' more significant achievements included passing on to the Romans the alphabet, which they had themselves borrowed from the Greek world, and acquainting them with certain Hellenic religious and art forms.

Etruscan influence on Rome

It should not be concluded, however, that Rome owed a great debt to its 6th-century Etruscan masters. On the contrary, the limits of their influence should be emphasized. Although the Etruscans physically transformed and reorganized Rome, the city on the whole withstood any profound exertion of Etruscan influence. The Latin tongue was little influenced by the Etruscan language, which had a quite different structure. In addition, once the Etruscan occupation had ended, Etruscans and Romans followed widely divergent paths. Rome had changed considerably in outer appearance, but its people, who, after a century of foreign domination, patiently and stubbornly continued their progress and history, remained the same.

The development of Roman institutions. At the very basis of the Roman city were the gentes, groups composed of all those linked by a common ancestor, who gave the gens its name and character. Each gens had its own cults and traditions that stubbornly withstood the ravages of time. In addition to those citizens bound together by the blood of the gentes, the city included a varying number of *clientes,* men who owed allegiance to patrons in return for aid and assistance. The gens comprised subdivisions, the *familiae,* the more limited but fundamental units of Roman society. They were grouped around the head of the family, the *paterfamilias,* whose authority is often demonstrated in Roman history by striking examples.

As far back as the period of monarchy, a privileged class, the patriciate, gradually began to form, made up of members of the privileged gentes. Opposed to it was the plebs, which was excluded from the gens organization. The plebeians had neither a cult nor common ancestors. They initially formed an unorganized class made up of former clients of vanished gentes and foreigners who had settled along the Tiber as traders or artisans. Lacking civil, political, or religious rights, the plebs initially also had no civic obligations and was liable to neither military service nor taxation. Although the reform attributed to Servius Tullius is full of dating errors, the Etruscan monarchy, following the principle of organizing the city according to census figures, was responsible for re-enlisting the plebs into the army.

The intense religious life of ancient Rome had achieved in remote times the essential organization it was to maintain for centuries. Greek and Etrurian influences were assimilated but did not profoundly alter the traditional religious structure. Several priesthoods of varying importance were charged with administering the *jus divinum* ("divine law"), keeping order within the city, and maintaining good relations with the gods. In the period of the monarchy, the king possessed religious authority in addition to his political, judicial, and military roles. Under the republic, this authority was passed on to a priest appointed for life, the *rex sacrorum,* who resided in the Regia in the Roman Forum. But the real religious power lay elsewhere. *Flamines* ("priests"), of which there were three major and 12 minor in the classical epoch, were assigned to the cult of individual deities. The *flamines* of Jupiter, Mars, and Quirinus were of remote origin and guaranteed the benevolent presence of the ancient triad of gods. The au-

The priesthoods

gurs were a body of theologians responsible for preserving and applying the rules pertaining to the observation and interpretation of the auspices, divine omens revealed by, among other things, the flight of birds. The pontiffs, particularly the *pontifex maximus,* held wide powers. They were the trustees of sacred knowledge and saw that it was strictly applied. They supervised the venerable college of the Vestals, virgin priestesses appointed for 30 years, who were charged with keeping the sacred fire of the city in the Temple of Vesta. There were also ancient brotherhoods, such as the Arvales and Salii, who assured the survival of ancient magic rituals.

The political and military power of the kings was absolute. This power, called the *imperium,* developed out of the right of military command but also had a religious dimension. The ceremony of the king's investiture was, in fact, controlled by a prior examination of the auspices through the offices of the augur. According to annals, each accession to the throne was normally contingent on a popular vote and the Senate's confirmation. What is certain is that the Roman monarchy was not hereditary and that it functioned for at least two and one-half centuries.

Early political divisions

According to legend, the kings, beginning with Romulus, organized the Roman people politically. They supposedly were divided into three tribes and 30 *curiae.* The primitive tribes may have corresponded to various ethnic elements: Latins, Sabines, and Etruscans. Each tribe was divided into ten *curiae,* which resembled the gentes and were local in character. Each *curia* was headed by a *curio* and had common meals and sacrifices. Together they formed the assembly called the Comitia Curiata, which elected the king and possessed a certain number of legislative and judicial rights. Such was the general structure that controlled the life of Rome during the early centuries of its existence. Simultaneously stable and flexible, it made possible the progressive development of a town that was self-governing and at the same time open to constructive influences from abroad. The cardinal Roman virtues, which were Rome's most steadfast support and the guarantor of its survival, had already asserted themselves. Rome was characterized by a profound patriotism that consistently subordinated the individual to the group and always resolved the conflicts of duty to the benefit of the state. The deep sense of the permanence of the state assured a continuing respect for ancestral custom, although it did not entail the rejection of innovation. But this continuity could be assured only by the exercise of a high level of morality, which gave first place to the civic virtues of unwavering piety and good faith. *Virtus, pietas, fides:* these guaranteed the survival and growth of Rome.

Early centuries of the Roman Republic (6th century BC–264 BC). *The overthrow of the monarchy.* The annals give a coherent and highly colourful account of the fall of the monarchy and the establishment of the republic. According to this account, the tyranny of Tarquinius Superbus became hateful to the Romans, who rose under the leadership of Lucius Junius Brutus after the rape of Lucretia. The Tarquins were banished from Rome; the popular assembly, grouped by centuries, named the first two magistrates in 509, Brutus and Tarquinius Collatinus. The king of Clusium, Lars Porsena, laid siege to Rome in an attempt to restore Tarquinius Superbus to power; but, according to legend, the heroism of Horatius Cocles and Mucius Scaevola forced him to retreat.

Dedication of the temple of Jupiter Capitolinus

The dedication of the temple of Jupiter Capitolinus by Marcus Horatius, who replaced Spurius Lucretius, Brutus' successor, as magistrate, also took place in 509 after the exile of Tarquinius Superbus and his family. In the same year was allegedly begun the custom of drawing up the list of eponymous magistrates, the consular *fasti,* the first and foremost document of Roman history.

Analysis of the various elements of the tradition shows that the annalists, for patriotic reasons, more or less consciously telescoped events that occurred over several years into that *annus mirabilis,* 509. The reliable basis for the chronology of the period is the dedication of the Capitoline temple of Jupiter, Juno, and Minerva, which took place in 509 BC. At that time began the annual rite of driving a nail into the wall of Minerva's cella in the Capitoline sanctuary. By counting the number of nails the true date of the dedication was easily obtained. But the Etruscan departure from Rome and the subsequent founding of a republic freed from foreign occupation has to be placed later, even as late as 470 BC. Archaeological data and the persistence of religious ceremonies show that until that time there was no break in Roman life. The Etruscans abandoned Rome and Latium only after they had been defeated by the Greeks in the waters of Cumae (474 BC), when they were already on the decline. The annals placed their departure in 509 BC because patriotism required it. The Capitoline temple, which was to become the symbol and centre of Rome and its empire, could not remain what it had in fact been, a purely Etruscan achievement. Its dedication at least had to be Roman. By arbitrarily placing the expulsion of the Tarquins in 509, the annalists could ascribe the glory of the dedication to Horatius, a true Roman.

Later, the expulsion of the Tarquins was regularly celebrated as the end of a monarchy that stood for the domination of an individual and the servitude of the citizens. In the folklore of later Romans it was at this point that *dominatio* and *servitus* vanished from Roman soil and the state became the property of its legitimate owner, the Roman people, who exercised power. Such was the foundation of Roman liberty (or *libertas*), the prestige of which remained such that even Augustus, the first emperor, improperly laid claim to it by proclaiming himself *libertatis vindex* ("saviour of liberty").

The chief magistracy and the dictatorship. With the end of the monarchy, a republican constitution gradually began to take form. The Romans' pragmatism led them to create a whole series of powers designed to balance one another while providing the maximum effectiveness to the essential proceedings of the state. Power was divided among the Senate, the popular assemblies, and magistrates of varying degrees of importance. What gradually took shape was an essentially oligarchic republican constitution in which a governing class, the *nobilitas,* consistently held the foremost position.

The king's sovereign executive power appears to have passed initially to a magistrate with the title of *praetor maximus,* who no doubt headed a body of *praetores,* who had perhaps been the immediate officers of the monarch. It was a *praetor maximus,* Marcus Horatius, who in 509 dedicated the Capitoline temple.

The consuls

The *praetor maximus,* however, soon gave way to a collegiate system in which two elected consuls shared the *imperium* for a single year. The principles of collegiality and of limitation of the term of office were characteristic of the republican constitution and effectively prevented the reappearance of tyrannical and absolute authority. It thus appeared safe to grant royal insignia to the magistrates and to have them preceded by 12 lictors bearing the fasces and the axe, hidden to Rome but visible on the other side of the *pomoerium,* the trench that defined the city limits.

Apparently, the plebeians initially were not excluded from the consulate, but the patriciate quickly closed ranks and made it extremely difficult for anyone not of its order to accede to the chief magistracy. In the middle of the 5th century, pressure from the plebs caused the patricians to create the military tribunate with magistral authority, a kind of promagistracy open to plebeians (see below). The consulate, however, remained open only to patricians until the mid-4th century BC, when a law was passed requiring that one of the two consuls be a plebeian. By that time the ancient distinction between the patriciate and the plebs had become blurred, and a new nobility occupied the forefront of the political scene.

During times of national emergency, the dual consulship was abandoned and a dictator established in their place. The dictator was, in fact, appointed by one of the consuls on the initiative of the Senate, and he was free of the restrictions imposed on the other magistrates. The dictator's *imperium* was unlimited but was entrusted to him for six months only, after which he surrendered his authority, as in the legend of Cincinnatus, who was appointed dictator

in the war against the Aequi and afterward returned to his farm. The dictator was backed up by a cavalry commander (or *magister equitum*) of his own choosing, who had to resign at the same time as the dictator. This special form of magistracy was seldom used. It no doubt arose from an early institution of the Latin towns, which reserved special dictatorships for carrying out certain religious rites. The dictator disappeared from Rome after the Second Punic War and reappeared only during the struggle of the 1st century BC among ambitious men who used it to fight their way to personal power.

Other magistracies. Initially, the *imperium* was exercised only by the consuls, but in 366 BC, according to tradition, a praetor was created who also possessed *imperium* but of an order inferior to that of the consuls (*imperium minus*). The praetor was given the responsibility for the administration of justice in the city. Beginning in the second half of the 3rd century, the growth of Rome and its conquests entailed the creation of other praetors who were charged with administering the law in the provinces.

The praetorship

Tradition credits King Servius Tullius with having taken the first census of the citizens and their property. Under the republic, a magistracy, important but *sine imperio,* fulfilled this function. This was the censorship, created, according to tradition, in 433. Two censors were elected every five years, but at the end of 18 months, after they had completed their task, they resigned their posts. They made a census of persons and property and were able to bring disgrace upon a citizen for bad conduct. They thus exercised a true *cura morum,* and this concern for morality made them important and respected figures.

As administration became increasingly complex, more officials were created to aid all of the high magistrates. The quaestors specialized in financial questions and were the custodians of the public coffers. The college of four aediles (two patricians and two plebeians) was charged with maintaining public order in the town and provisioning it, along with the upkeep of public monuments and the organization of games.

Judicial institutions. From remotest times the Romans displayed a concern for legal rules, the religious character of which was striking. According to tradition, a religious brotherhood of 20 members (the *fetiales*) appeared as far back as the 8th century BC, in possession of the laws of warfare. War could not be declared nor treaties concluded without applying strict rules, of which they were the trustees. Rome recognized neither the rule of the strongest nor the right of the conqueror over the vanquished in warfare. The Romans' respect for legal obligations permitted them to conduct only a war that was *pium justumque* ("pious and just"), so that the gods would aid the Roman army to victory. Piety, good faith, and justice were interrelated and, in the mind of the Romans, were needed to guide their struggles and justify their conquests.

Law of the Twelve Tables

In the middle of the 5th century, the law of the Twelve Tables was framed and displayed in the Forum on 12 bronze tablets. It was Rome's first written code and was of extreme importance because it was viewed as the source of all law, public as well as private. The schoolchildren of Cicero's time still learned it by heart. The code survives only in fragments, but enough remains to enable modern historians to grasp its spirit.

The code seems to have been the creation of a special committee. There is an interruption in the consular *fasti* for 451 and 450, when the consuls were replaced by a board of ten men called decemvirs, who ruled with consular authority and who were charged with putting the laws into written form. The first college of decemvirs was appointed for one year and was replaced by a second college that was to complete the task; but the second college tried to retain power illegally. Military defeats and scandals in Roman life brought about its downfall and the consuls' return to power. During their brief rule, however, the decemvirs had sufficient time to fulfill their task of drafting a written code, to which anyone might refer. Until then the law had remained purely oral, and the plebs accused the magistrates of applying it unfairly. According to the ancient writers and from the fragments of extant laws, the code of the Twelve Tables seems to be the heir of a remote past, for many of its provisions present an essentially religious and sacred character. The code, however, also displays the distinguishing feature of judicial law: the assurance to every citizen of the exercise of his rights.

The Senate. In the political life of republican Rome, the Senate constituted a third element, in addition to the popular assemblies and magistracies. Its prestige was great and its influence of continuing importance. The republican Senate was heir to the Senate of the royal epoch, which, according to tradition, was created by Romulus and brought together the most influential of the *patres familiarium.* The royal Senate had been responsible for nominating the king and was, no doubt, his constant adviser, already enjoying an *auctoritas,* an effective power of moral force with religious overtones, which lent value to its acts and gave them unusual weight in the citizens' eyes. After the fall of the monarchy, the Senate was presided over by the acting consul and represented the most stable political authority of the republic.

The Senate was originally composed of heads of the patrician gentes but soon came to include all former magistrates of the city, whose names were entered on the roll and who thus became *conscripti.* This probably explains the name borne by the Senators under the republic—*patres conscripti.* This term denoted two groups of different origin and should be understood as having arisen from the formulation *patres et conscripti.* At first senators were appointed by the consuls, and later by the censors, so that changes in the Senate's composition occurred only every five years, when new censors were elected. Senators were seated for life, and their hierarchy corresponded to that of the magistracies that they had previously held.

The Senate's *auctoritas*

The Senate's *auctoritas* manifested itself in all areas. In the early years of the republic, the Senate approved the laws after they had been voted upon in the assemblies, but from the second half of the 4th century, Senate action no longer followed but, instead, preceded the popular vote. Senatorial assent was succeeded by the right of initiative in proposing laws. The Senate expressed its opinions in the form of a *senatus consultum* (a decree of the Senate), the phrasing of which followed an unvarying scheme. Although it was not obligatory, the magistrates were aware that they should abide by the opinion of the wisest and most respected men of the city. In time, the Senate's opinions gained the force of law.

In the later republic, the Senate also performed the important function of assigning magistrates to govern the provinces. Since the political career of a Roman statesman usually depended upon obtaining a lucrative provincial appointment, this became of great significance in the power struggle of the 1st century BC.

The Senate also had great authority in foreign affairs. The ambassadors of foreign powers presented themselves before it, and it chose the envoys to be sent abroad. With respect to religion, it upheld the ancestral rites and ceremonies, accepting foreign divinities when the need was felt and securing Rome against innovations dangerous to the *mos majorum.* Above all, it had supreme control over public finances since it managed the state treasury (the *aerarium Saturni*) and made final decisions on expenditures and taxation.

In the early centuries of the republic, the Senate thus had supreme influence, due to custom and to the respect in which it was held by the magistrates and popular assemblies. The Roman constitution, which appeared to be based on a combination of forces, was in fact essentially oligarchic and plutocratic. The senators, who were proprietors of great estates, were thrifty, competent administrators. Their real difficulties arose with conquest, which gave them not only a town and a country to govern but a true empire to administer as well.

Plebeian institutions. In a patrician state that wished to exclude all those who had not issued from its ranks, the plebeians, forced into the background, put all of their efforts into forcing recognition of their existence and rights. The first two centuries of the republic were taken up by their unceasingly renewed attempts to gain the political position that they claimed and by the progressive disinte-

gration of the privileges to which the patricians clung tenaciously. The struggle took place on the economic as well as on the civic levels. The plebeians wanted the expanses of the public lands (the *ager publicus*), over which the great proprietors were extending their influence, opened to them. The public domain was expanding continually as a result of the gradual conquests of the Roman armies.

Struggle over public land

The long conflict between patriciate and plebs went through various changes, described by the annalists in accounts that are often unreliable and contain errors of dating. What is certain is that in the first decades of the republic the plebs was able to organize itself and form a state within the state. The unity of the republic was ultimately maintained, but not until serious crises had threatened the very foundations of its existence.

The process of organization and defense of the plebeians was apparently as follows: around the beginning of the republic, 493–92, according to tradition, the plebs, returning armed from a military campaign, seceded to the Sacred Mount and forced the Senate to grant to it magistrates charged with defending the plebs under all circumstances. The tribunes of the plebs were inviolable and sacrosanct and could bring aid (*jus auxilii*) to plebeians threatened by the possible excesses of the executive power. They also had a right of veto, through which they could prevent the consuls, Senate, or assembly from making a decision, and they could intercede to prevent the carrying out of a decision if they judged it to be unfairly detrimental to the plebs. The tribunes' power, which they exerted only within the city, operated within the sphere of magic and religion. In the face of the regular Roman authorities, it appeared that the plebs could be protected only by figures from its own ranks, who were effectively protected by divine power and who themselves enjoyed a sacred power from the depths of antiquity. Immediately after the tribunates were established, the consul Spurius Cassius dedicated a tripartite temple at the foot of the Aventine, which was constructed on the model of the Capitoline sanctuary but devoted to Ceres, Liber, and Libera. This was to be the appointed sanctuary of the plebs. Ceres was the protectress of agriculture and kept the deadly famine away from the urban population. The couple Liber and Libera presided over the birth and growth of animals and people. Permanently established on the Aventine, this triad of fertility protected the plebs, whose land and herds guaranteed only an uncertain subsistence, and, to some extent, counterbalanced the Capitoline triad, the great protector of the Roman state and its regular authorities, who were patricians. The temple on the Aventine held the plebeian archives and, in the later republic, became the centre for wheat distributions to the plebs. Aediles were created to protect the temple, guard the archives, and keep order in the markets.

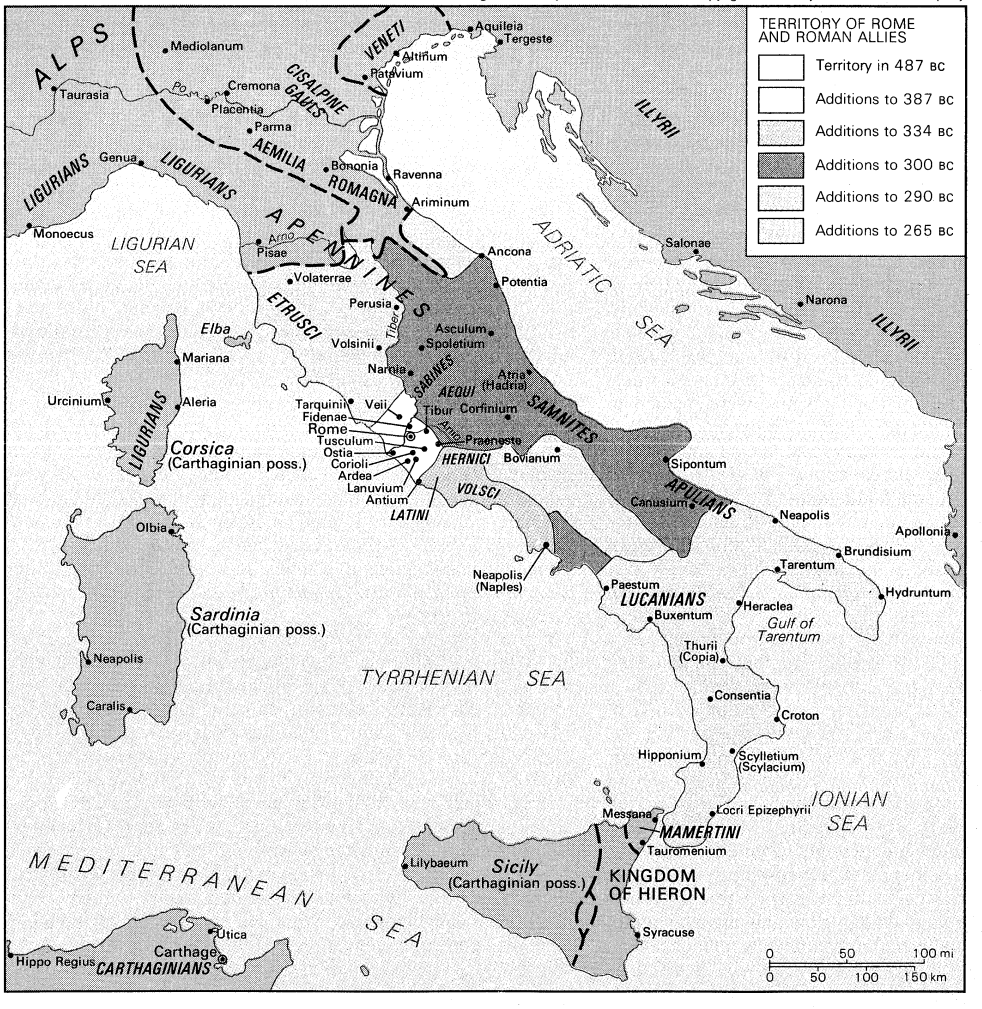
Plebeian religious institutions

Completing the organization of the plebeian class was the assembly of the plebs, the Concilium Plebis, which was organized according to the citizens' domiciles instead of by property qualifications, as in the Comitia Centuriata.

The state was thus apparently divided in half. Rome was wise and fortunate in its ability, through successive measures, to bring the opposing orders to an understanding. A law of 456, which determined the sale of the lands on the Aventine, signalled the first extension of the *ager publicus*.

The expansion of Rome in Italy. *Rome and its Latin neighbours.* Roman historians provide a detailed picture of Rome's foreign policy and its conquests during the early centuries of the republic. There is splendour and grandeur in that account, which often takes on an epic rhythm, in which Rome, engaged with many powerful

From A. Boak and W. Sinnigen, *A History of Rome to A.D. 565*; Copyright © 1965 by The Macmillan Company



Roman expansion in Italy 487–265 BC.

adversaries, fells them one by one, relying on allies, of course, but bearing the burden of the battles and struggles itself. There are many more or less conscious errors of dating in these highly colourful pages, whose coherence and frequent beauty filled the Romans of the classic era with satisfaction and pride in their ancestors. One important error of perspective makes Rome a capital among the other towns of central Italy as early as the 5th century BC and gives the city a preeminence over neighbouring cities that it certainly did not have. Modern historical research makes it possible to take a more realistic view of the situation.

The Etruscan departure from Latium to a certain extent weakened Rome, for the Etruscans had fortified Rome as a base for their expeditions to the south of the peninsula.

The Latin League

Rome joined the Latin League and had to adopt its policy since it certainly was incapable of imposing its will on it. The entire 5th century and the first part of the 4th was the era of the *concilium Latinorum.* Decisions were made in common, and in time of war the Romans, along with everyone else, placed themselves under the command of the dictator elected by the league.

These common perils did not prevent Rome from breaking with its Latin neighbours. Rome's victory at Lake Regillus in 499 BC and the treaty of alliance with the Latin towns—known as the Foedus Cassianum from the name of the consul, Spurius Cassius, who signed it, and ascribed by tradition to the year 493—must be put back into the framework of the Etruscan occupation, which had not yet ended. The great conflicts between republican Rome and its Latin rivals for preeminence over Latium have to be placed later. The major problem during the 5th century was curbing the peoples of the central Apennines, who were multiplying their incursions into the western plains. These were the Sabines, the ancient enemies of monarchic Rome; the Aequi, whose territory extended from the Anio to Praeneste; and their neighbours, the Hernici and the Volsci, who occupied the Lepini mountains and directed their raids toward Ardea and Antium. Somehow the pressure of the Sabellian peoples was contained, and Rome opened a passage into the Volscian country by occupying Corioli in 444.

Rome found itself at the same time engaged in a long conflict with the nearest of the Etruscan cities, Veii, located on the right bank of the Tiber. Veii was a powerful and formidable rival, and the struggle was to be long and to undergo many changes before Rome finally triumphed over its opponent.

Fidenae, situated on the Tiber a short distance above Rome, was a port of great importance that blocked the salt route going from the sea to the mountains, used by the Etruscans. It was the stake in the successive wars between Rome and Veii, which firmly held Fidenae. Rome did not definitely take possession of it until 426 BC.

The siege of Veii

Then, according to tradition, in 405 began the siege of Veii itself, which lasted ten years (like the siege of Troy) and ended in the city's collapse under the blows of the dictator Furius Camillus. Livy paints an epic picture of this siege and its dramatic outcome. The religious notes that sprinkle his narrative certainly derive in part from traditions that are unmistakably Etruscan. Fate and the oracles had announced the impending end of Veii, abandoned by the gods themselves. Camillus, for his part, was fate's moving spirit. He sent an envoy to the Delphic Apollo and promised the tithe of the spoils if he were successful. He was assured of support, but he was loath to attack the town without appealing to Juno, the goddess who protected it. He therefore practiced the ancient rite of *evocatio,* inviting the deity to leave the besieged town and come to live in Rome, where she would be received with honour. Juno accepted and the town was stormed and captured, after which the statue of the cult left the ruins of Veii and arrived on the Aventine, where it was given a temple suitable to Juno's dignity.

The Gallic invasion and further conquests. The capture of Veii was followed by a catastrophe for Rome that lived on in its history as a disastrous event, the Celtic invasion and the capture of the city by the Gauls. Book V of Livy and the pages dedicated by Polybius to the Celtic world make it possible to portrary vividly the vicissitudes of this dramatic period. At the end of the 6th and the beginning of the 5th century, the first bands of invaders began to arrive in northern Italy. In the 4th century the trickle became a flood, and the barbarian hordes penetrated the peninsula in force. According to the annalists, the Gauls allegedly defeated the Etruscan army near Clusium and continued southward toward Rome and crushed the Roman forces on the banks of the River Allia, to the north of the city. About 390 (the traditional date) they succeeded in routing the Roman army, which was dismayed by the appearance and mode of combat of its new adversaries. This day remained in the memory of the Romans a *dies religiosus* (an "accursed day"). Caere, Rome's Etruscan neighbour, received the sacred objects (the *sacra*) rescued by the Vestals from the pillage and ruin, after which a common religious atmosphere seemed to envelop the two cities. Once Rome was out of danger, its gratitude to the Caerites prompted it to grant them public hospitality, and it soon gave them citizenship without suffrage (*civitas sine suffragio*). After the drama of the occupation of Rome, which engendered an acute feeling of peril revived by Livy with a sort of tragic grandeur, the Gauls were summoned back to the north. Other Celtic invasions followed during the 4th century, but none gave rise to the terror of the first assault.

Since the Latin confederation had abandoned Rome to its own devices, the Romans had confronted the invaders virtually alone. There ensued a period of anarchy in Latium, and Rome took advantage of it to subjugate the disordered Latin towns. A temporary agreement between Rome and the Latin towns was made in 358 BC.

The Samnite Wars

Through a complicated shifting of alliances, Rome's first war against the Samnites in 343 BC brought about a decisive war between Rome and the Latins, which resulted in the dissolution of the Latin League and the end of the common policy undertaken, at least in theory, by Rome and its neighbours. The latter lost the advantages gained by confederacy and were assigned various statuses, becoming allied cities, like Tibur and Praeneste, or municipia, like Tusculum and Lavinium. Henceforth, Rome dominated Latium.

Campania was opened to Roman expansion, and Rome multiplied friendly contacts with the Campanian privileged classes and with such Greek coastal cities as Naples. The Oscan and Greek populations were constantly threatened by the incursions of the Samnites, a martial mountain people. The Second Samnite War, which lasted from 326 to 304, and the third, which extended between 298 and 290, were necessary to subjugate the powerful stronghold of Samnium. The Samnites, who faithfully obeyed their federal magistrates, were moved by a fierce and bellicose spirit. Well-armed and good tacticians, they inflicted heavy losses on the Romans. In 321 they trapped the Roman army in the closed valley of Caudine Forks between Capua and Beneventum; the two consuls, who survived, accepted an agreement that was harsh for Rome, and the entire army surrendered. Rome did not ratify the Pax Caudina, but it had suffered a deep humiliation. Its tenacity enabled it to regain the advantage, and Bovianum, one of the Samnite capitals, was stormed in 306, and the peace signed. At the same time, Rome pressed into central Etruria, which was living out its last years of independence. The Third Samnite War broke out in 298, in which the Etruscans, Gauls, and Samnites united against the increasing threat of the Romans. In 295 the consuls Fabius Rullianus and Decius Mus won a great victory at Sentinum, north of the Umbria, over the Samnito-Gallic coalition, in which Decius, according to legend, heroically sacrificed his own life. The Gallic tide was stemmed, and the Ager Gallicus Romanized. The conquest of Etruria was completed with the fall of the last independent Etruscan town, Volsinii, in 265 BC.

The southward expansion of Rome, which was still on good terms with Carthage, transformed the republic into a Mediterranean power. The Romans came into contact with peoples such as the Lucanians and the Apulians, over whom they extended their influence, which engendered the hostility of the most powerful city of Magna Graecia,

Wars with Pyrrhus

Tarentum. When a Roman garrison and ships appeared in areas of Tarentine authority, Tarentum, sensing danger, decided to put a stop to it and appealed to Pyrrhus, the king of Epirus, to bring the insolent Romans to their senses. Pyrrhus thought he could unify the Hellenic cities of southern Italy and the native populations of the interior. Although this ambition miscarried, his expeditionary corps, using formidable combat elephants, defeated the legions on several occasions between 280 and 276 BC. From Carthage, then friendly to Rome, he won back the Greek cities of Sicily that had fallen under Punic hegemony. But Pyrrhus' victories were short-lived. He returned to Epirus, and Tarentum had to concede defeat in 271. Rome accorded Tarentum the status of an allied city.

The Roman mastery of Italy. Thus, by the beginning of the 3rd century BC, Rome had become mistress of Italy and was rapidly becoming a Mediterranean power. Tarentum had bequeathed to Rome its hegemony in the commerce of the Adriatic. The Greek world for the first time turned its gaze toward this latest arrival on the international chessboard, this city of the west whose dynamism asserted itself through a series of conquests that neither difficulties nor defeats had been able to check. In 306 the agreement with the Carthaginians was renewed once more, the last time before the outbreak of the hostilities that were to lead to great upheavals. This last treaty forbade all Carthaginian intervention in Italy and all Roman intervention in Sicily. Forty years later, however, Rome interfered in Sicilian affairs, and the former allies found themselves in opposition.

The Roman concept of imperialism

In its conquest of Italy, Rome had given proof of a flexible diplomacy, which treated the vanquished honorably and gave them a place, and rights that varied, in a country gradually being unified under Roman law. Roman imperialism sought constantly to justify itself morally, and the demands of Roman good faith (*fides Romana*) were not just empty words, even though such honesty was not always observed. Through *deditio,* the vanquished relied on the *fides* of Rome, which claimed to use the new power conferred on it for the good of all, in the best way possible. In point of fact, the defeated city usually lost its autonomy and incurred various obligations to Rome, so that after the dissolution of the Latin League, the status of Rome's former allies was diverse. Lanuvium and Tusculum, for example, had citizenship without suffrage (*civitas sine suffragio*), while Tibur and Praeneste remained allied cities (*civitates foederatae*), which did not prevent them from losing part of their territory. Rome also sent new colonists into the conquered regions, and the Roman colonies spread throughout the ever-expanding *ager publicus* were like extensions of the metropolis in a more or less distant land. The population of the Latin colonies was composed largely of proletarians who cleared the conquered soil and brought it under cultivation. Because of this vast extension of Rome's national territory, economic development of the city was rapid. The city soon had need of a currency of its own for its expanding and intensified commercial relations. The war with Pyrrhus gave rise to the first Roman silver coinage.

On the eve of the First Punic War, Rome had become a commercial city participating in the great economic currents of the Hellenistic world. The wars against Carthage forced Rome to become a naval power. Rome could then operate in the east and undertake the expeditions and conquests beyond the seas that ultimately would lead to the extension of its hegemony throughout the Mediterranean world.

(R.Bl.)

THE MIDDLE REPUBLIC (264–133 BC)

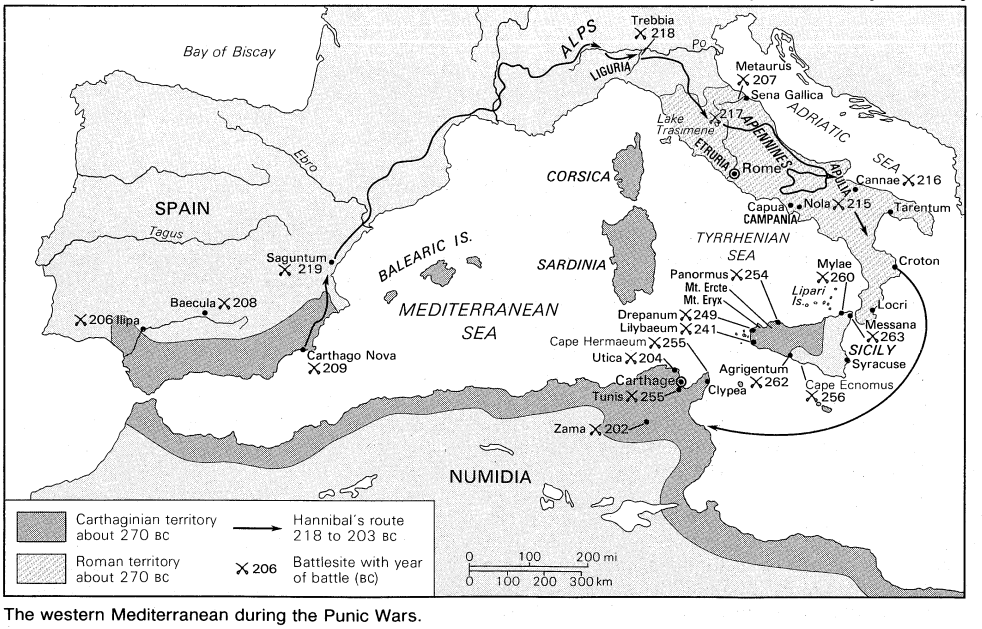
The first two Punic Wars. Rome's desire to safeguard the Italian peninsula against foreign interference made it necessary to prevent the neighbouring island of Sicily from falling into the hands of a strong and expansive power. Carthage, on the other hand, had long been anxious to conquer Sicily and to complete the chain of island posts by which it controlled the western Mediterranean from its base in North Africa.

First Punic War (264–241 BC). The proximate cause of the first outbreak was a crisis in the city of Messana (Messina), commanding the straits between Italy and Sicily. A band of Campanian mercenaries, the Mamertini, who had forcibly established themselves within the town and were being hard-pressed in 264 by Hieron II of Syracuse, applied for help to both Rome and Carthage. The Carthaginians, arriving first, occupied Messana and effected a reconciliation with Hieron. The Roman commander, nevertheless, persisted in throwing troops into the city and by seizing the Carthaginian admiral during a parley induced him to withdraw. This aggression involved Rome in war with Carthage and Syracuse.

Operations began with a joint attack upon Messana, which the Romans easily repelled. In 263 the Romans advanced with a considerable force into Hieron's territory and induced him to seek peace and alliance with them. In 262 they besieged and captured the enemy's base at Agrigentum. But they made little impression upon the Carthaginian fortresses in the west of the island and upon the towns of the interior.

In 260 the Romans built their first large fleet of standard battleships. At Mylae (Milazzo), off the north Sicilian coast, their admiral Gaius Duilius defeated a Carthaginian

Adapted from *Westermann Grosser Atlas zur Weltgeschichte;* Georg Westermann Verlag, Braunschweig



The western Mediterranean during the Punic Wars.

squadron of superior manoeuvring capacity by grappling and boarding. This left Rome free to land a force on Corsica (259) and expel the Carthaginians but did not suffice to loosen their grasp on Sicily. A large Roman fleet sailed out in 256, repelled the entire Carthaginian fleet off Cape Ecnomus (near modern Licata), and established a fortified camp on African soil at Clypea (Kélibia in Tunisia). The Carthaginians, whose citizen levy was utterly disorganized, could neither keep the field against the invaders nor prevent their subjects from revolting. After one campaign they were ready to sue for peace, but the terms that the Roman commander M. Atilius Regulus offered were intolerably harsh. Accordingly, the Carthaginians equipped a new army in which, by the advice of a Greek captain of mercenaries named Xanthippus, cavalry and elephants formed the strongest arm. In 255, under Xanthippus' command, they offered battle to Regulus, who had taken up position with an inadequate force near Tunis, outmanoeuvred him, and destroyed the bulk of his army. A second Roman fleet, which reached Africa after defeating the full Carthaginian fleet off Cape Hermaeum (Cape Bon), withdrew all the remaining troops.

The campaign against Sicily

The Romans now directed their efforts once more against Sicily. In 254 they captured the important fortress of Panormus (Palermo), but when Carthage threw reinforcements into the island the war again came to a standstill. In 251 or 250 the Roman general Caecilus Metellus at last brought about a pitched battle near Panormus, in which the enemy's force was effectively crippled. This victory was followed by an investment of the chief Punic base at Lilybaeum (Marsala), together with Drepanum (Trapani), by land and sea. The besiegers met with a gallant resistance and in 249 were compelled to withdraw by the loss of their fleet in a surprise attack upon Drepanum, in which the admiral P. Claudius Pulcher was repulsed with a loss of 93 ships. This was the Romans' only naval defeat in the war, but their fleet had suffered a series of grievous losses by storm, and now it was so reduced that the attack upon Sicily had to be suspended. At the same time, the Carthaginians, who felt no less severely the financial strain of the prolonged struggle, reduced their forces and made no attempt to deliver a counterattack. The only noteworthy feature of the ensuing campaigns is the skillful guerrilla war waged by a new Carthaginian commander, Hamilcar Barca, from his strong positions on Mt. Ercte (247/246–244) and Mt. Eryx (modern Erice; 244–242) in western Sicily, by which he effectually screened Lilybaeum from any attempt on it by the Roman land army.

In 242 Rome resumed operations at sea. By a magnificent effort on the part of private citizens, a fleet of 200 warships was equipped and sent out to renew the blockade of Lilybaeum. The Carthaginians hastily collected a relief force but in a battle fought off the Aegates, or Aegusae (Aegadian) Islands, west of Drepana, their fleet was caught at a disadvantage and mostly sunk or captured (March 10, 241). This victory, by giving the Romans undisputed command of the sea, rendered certain the ultimate fall of the Punic strongholds in Sicily. The Carthaginians accordingly opened negotations and consented to a peace by which they ceded Sicily and the Lipari Islands to Rome and paid an indemnity of 3,200 talents.

The interval between the First and Second wars (241–218 BC). The loss of naval supremacy not only deprived the Carthaginians of their predominance in the western Mediterranean but exposed their overseas empire to disintegration under renewed attacks by Rome. The temper of the Roman people was soon made manifest during a conflict that broke out between the Carthaginians and their discontented mercenaries. A gross breach of the treaty was perpetrated when a Roman force was sent to occupy Sardinia, whose insurgent garrison had offered to surrender the island (238). To the remonstrances of Carthage the Romans replied with a declaration of war and only withheld their attack upon the cession of Sardinia and Corsica and the payment of a further indemnity.

From this episode it became clear that Rome intended to use the victory to the utmost. To avoid complete humiliation Carthage had no resource but to humiliate its adversary. The recent complications of foreign and internal strife had indeed so weakened the Punic power that the prospect of renewing the war under favourable circumstances seemed remote. But the scheme of preparing for a fresh conflict found a worthy champion in Hamilcar Barca, who sought to compensate for the loss of Sicily by acquiring a dominion in Spain where Carthage might gain new wealth and form a fresh base of operations against Rome. Invested with an unrestricted foreign command, he spent the rest of his life in founding a Spanish empire (237–228). His work was continued by his son-in-law Hasdrubal and his son Hannibal, who was placed at the head of the army in 221. These conquests aroused the suspicions of Rome, which in a treaty with Hasdrubal confined the Carthaginians to the south of the Ebro. At some point Rome also entered into relations with Saguntum (Sagunto), a town on the east coast, south of the Ebro. In 219 Hannibal laid siege to Saguntum and carried the town in spite of a stubborn defense.

It has always been a debatable point whether or not his attack contravened the new treaty. The Romans certainly took this view and sent to Carthage to demand Hannibal's surrender. But his defiant policy was too popular to be disavowed; the Carthaginian council upheld Hannibal's action and drew upon itself a declaration of war.

Second Punic War (218–201 BC). It seemed as though the superiority of the Romans at sea must enable them to choose the field of battle. They decided to embark one army for Spain and another for Sicily and Africa. But before their preparations were complete, Hannibal began that series of operations by which he dictated the course of the war for the greater part of its duration. Realizing that, so long as the Romans commanded the resources of an undivided Italian confederacy, no foreign attack could beat them down beyond recovery, he conceived the plan of cutting off their supply of strength at the source by carrying the war into Italy and causing a disruption of the league. His chances of ever reaching Italy seemed small, for the sea was guarded by the Roman fleets and the land route was long and arduous.

Hannibal's strategy

But the very boldness of his enterprise contributed to its success; after a six months' march through Spain and Gaul and over the Alps, which the Romans were nowhere in time to oppose, Hannibal arrived (autumn 218) in the plain of the Po with 20,000 foot and 6,000 horses, the pick of his African and Spanish levies.

His further advance was here disputed by some Roman troops, but the superiority of the Carthaginian cavalry and the spread of insurrection among the Gaulish inhabitants forced the defenders to fall back upon the Apennines. At the end of the year the Roman army was reinforced by the division from Sicily and led out to battle on the banks of the Trebbia. Hannibal, by superior tactics, repelled the assailants with heavy loss and thus made his position in north Italy secure.

In 217 the campaign opened in Etruria, into which the invading army, largely reinforced by Gauls, penetrated by an unguarded pass. A rash pursuit by the Roman field force led to its being entrapped on the shore of Lake Trasimene (Trasimeno) and destroyed with a loss of at least 15,000 men. This catastrophe left Rome completely uncovered; but Hannibal, having resolved not to attack the capital before he could collect a more overwhelming force, directed his march toward the south of Italy, where he hoped to stir up the peoples who had formerly been the most stubborn enemies of Rome. The Italians, however, were slow everywhere to join the Carthaginians; and a new Roman army under the dictator Q. Fabius Maximus ("Cunctator") that, without ever daring to close with Hannibal, dogged his steps on his forays through Apulia and Campania, prevented his acquiring a permanent base of operations.

The Battle of Cannae

The eventful campaign of 216 was begun by a new, aggressive move on the part of Rome. An exceptionally strong field army, variously estimated at between 48,000 and 85,000 men, was sent to crush the Carthaginians in open battle. On a level plain near Cannae in Apulia, which Hannibal had chosen for his battleground, the Roman legions delivered their attack. Hannibal deliberately allowed his centre to be driven in by their superior numbers,

while Hasdrubal's cavalry wheeled round so as to take the enemy in flank and rear. The Romans, surrounded on all sides and so cramped that their superior numbers aggravated their flight, were practically annihilated; and the loss of citizens was perhaps greater than in any other defeat that befell the republic.

The moral effect of the battle was no less momentous. The south Italian peoples at least found courage to secede from Rome, the leaders of the movement being the people of Capua, at the time the second greatest town of Italy. Reinforcements were sent from Carthage, and several neutral powers prepared to throw their weight into the scale on Hannibal's behalf. At first sight it seems strange that the Battle of Cannae did not decide the war. But the great resources of Rome, though terribly reduced in respect to both men and money, were not yet exhausted. In north and central Italy the insurrection spread but little and could be sufficiently guarded against with small detachments. In the south, the Greek towns of the coast remained loyal, and the numerous Latin colonies continued to render important service by interrupting free communication between the rebels and detaining part of their forces.

In Rome itself the quarrels between the nobles and commons, which had previously unsettled Roman policy, gave way to a unanimity unparalleled in the annals of the republic. The guidance of operations was henceforth left to the Senate, which, by maintaining a persistent policy until the conflict was brought to a successful end, earned its greatest title to fame.

The subsequent campaigns of the war in Italy assumed a new character. Though the Romans contrived at times to raise 200,000 men, they could spare only a moderate force for field operations. Their generals, among whom the veterans Fabius and M. Claudius Marcellus frequently held the most important commands, rarely ventured to engage Hannibal in the open and contented themselves with observing him or skirmishing against his detachments. Hannibal, whose recent accessions of strength were largely discounted by the necessity of assigning troops to protect his new allies or secure their wavering loyalty, was still too weak to undertake a vigorous offensive. In the ensuing years the war resolved itself into a multiplicity of minor engagements, which need not be followed out in detail. In 216 and 215 the chief seat of war was Campania, where Hannibal, vainly attempting to establish himself on the coast, experienced a severe repulse at Nola.

In 214 the main Carthaginian force was transferred from Apulia in hopes of capturing Tarentum (Taranto). Though Croton and Locri Epizephyrii on the southern coast had fallen into his hands, Hannibal still lacked a suitable harbour by which he might have secured his overseas communications. For two years he watched in vain for an opportunity to surprise the town, while the Romans narrowed down the sphere of revolt in Campania and defeated other Carthaginian commanders.

In 213–212 the greater part of Tarentum and other cities of the southern seaboard at last came into Hannibal's power. But in 212 the Romans found themselves strong enough to place Capua under blockade. They severely defeated a Carthaginian relief force and could not be permanently dislodged even by Hannibal himself. In 211 Hannibal made a last effort to relieve his allies by a feint upon Rome itself, but the besiegers refused to be drawn away from their entrenchments, and eventually Capua was starved into surrender.

Its fall was a sign that no power could in the long run uphold a rival Italian coalition against Rome. After a year of desultory fighting, the Romans in 209 gained a further important success by recovering Tarentum. Though Hannibal still won isolated engagements, he was slowly being driven back into the extreme south of the peninsula.

The campaign of 207

In 207 the arrival of a fresh invading force produced a new crisis. Hasdrubal, who in 208–207 had marched overland from Spain, appeared in northern Italy with a force scarcely inferior to the army that his brother had brought in 218. After levying contingents of Gauls and Ligurians, he marched down the east coast with the object of joining his brother in central Italy for a direct attack upon Rome itself. By this time the steady drain of men and money was telling so severely upon the confederacy that some of the most loyal allies protested their inability to render further help. Yet by exerting a supreme effort the Romans raised their war establishment to the highest total yet attained and sent a strong field army against each Carthaginian leader.

The danger to Rome was chiefly averted by the prompt insight and enterprise of the consul Gaius Nero, who commanded the main army in the south. Having discovered that Hannibal would not advance beyond Apulia until his brother had established communications with him, Nero slipped away with part of his troops and arrived in time to reinforce his colleague Livius, whose force had recently got into touch with Hasdrubal near Sena Gallica (Senigallia).

The combined Roman army frustrated an attempt of Hasdrubal to elude it and forced him to fight on the banks of the Metaurus (Metauro). The battle was evenly contested until Nero, by a dexterous flanking movement, cut the enemy's retreat. Hasdrubal himself fell, and the bulk of his army was destroyed.

The campaign of 207 decided the war in Italy. Though Hannibal still maintained himself for some years in southern Italy, this was chiefly due to the exhaustion of Rome after the prodigious strain of past years and the consequent reduction of Roman forces. In 203 Italy was finally cleared of Carthaginian troops. Hannibal, in accordance with orders received from home, sailed back to Africa; and another expedition under his brother, Mago, which had sailed to Liguria in 205 and endeavoured to rouse the slumbering discontent of the people in Cisalpine Gaul and Etruria, was forced to withdraw.

Campaigns in Sicily and Spain. Concurrently with the great struggle in Italy, the Second Punic War was fought out on several other fields. It will suffice merely to allude to the First Macedonian War (214–205), which King Philip V commenced when the Roman power seemed to be breaking up after Cannae. This compelled the Romans to stretch their already severely strained resources still further by sending troops to Greece, but the diversions that Roman diplomacy provided for Philip in Greece, and the maintenance of a patrol squadron in the Adriatic Sea, prevented any effective cooperation on his part with Hannibal.

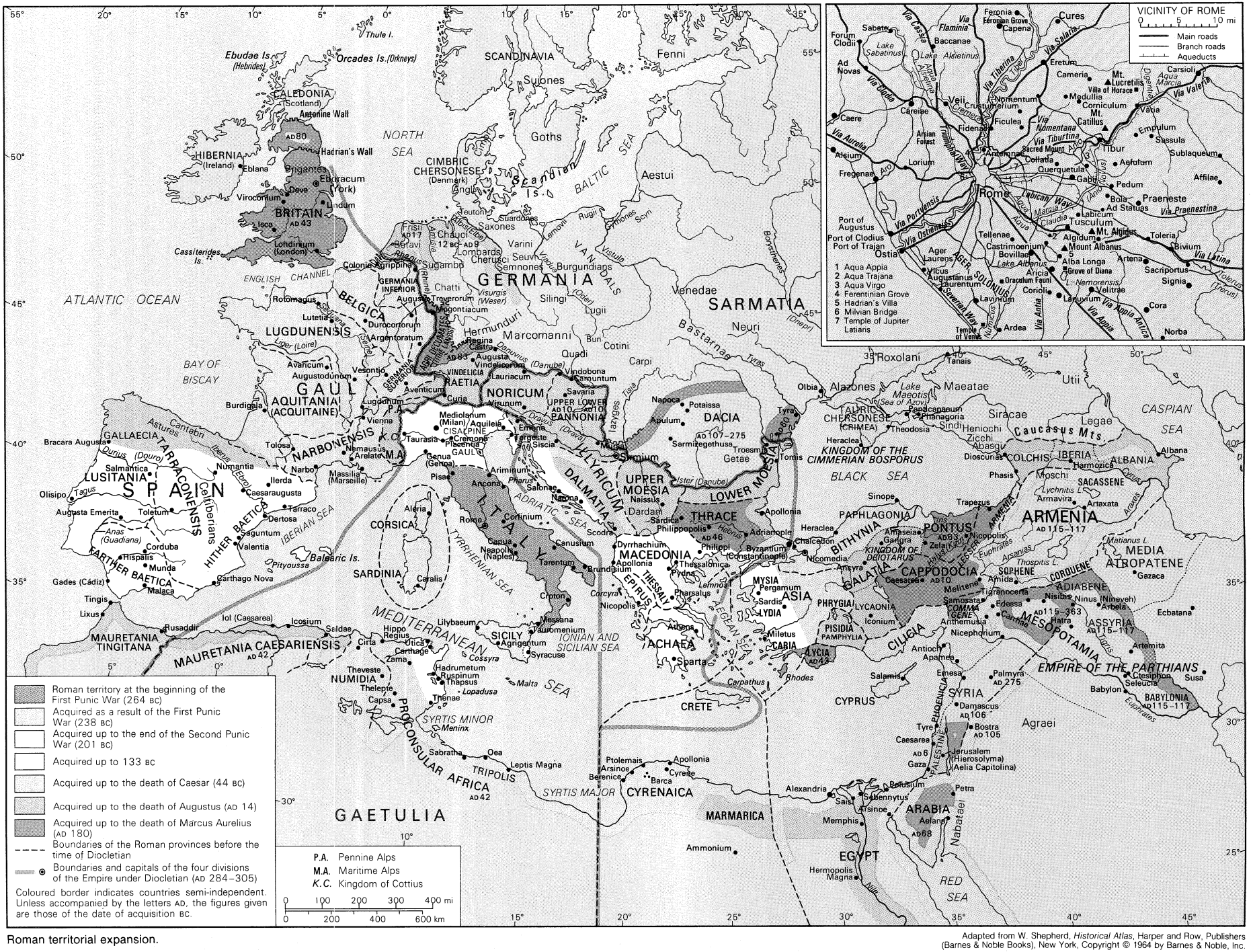
In view of the complete stagnation of agriculture in Italy the Romans had to look to Sardinia and Sicily for their food supply. Sardinia was attacked by Carthaginians in 215, but a small Roman force sufficed to repel the invasion. In Sicily a more serious conflict broke out. Some isolated attacks by Punic squadrons were easily frustrated by the strong Roman fleet. But in 215 internal complications arose. The death of Hieron II, Rome's steadfast friend, left the realm of Syracuse to his inexperienced grandson, Hieronymus. Flattered by the promises of Carthaginian emisaries, the young prince abruptly broke with the Romans, but before hostilities commenced he was assassinated. The Syracusan people now repudiated the monarchy and resumed their republican constitution; but, by threats of terrible punishment at the hands of Rome, they were led to cooperate with the Carthaginians.

Roman attack on Syracuse

The attacks of a Roman army and fleet under Marcellus, which speedily appeared before the town, were completely baffled by the mechanical contrivances of the Syracusan mathematician Archimedes (213). Meantime, the revolt against Rome spread in the interior, and a Carthaginian fleet gained control of towns on the south coast.

In 212 Marcellus at last broke through the defense of Syracuse and, in spite of the arrival of a Carthaginian relief force, mastered the whole town in 211. A guerrilla warfare followed in which the Carthaginians maintained the upper hand until in 210 they lost their base at Agrigentum. They were dislodged from their remaining positions, and by the end of the year Sicily was wholly under the power of Rome.

The conflict in Spain was second in importance only to the Italian war. From this country the Carthaginians drew large supplies of troops and money that might serve to reinforce Hannibal; hence it was in the interest of the Romans to challenge their enemy within Spain. Though



Roman territorial expansion.

Adapted from W. Shepherd, *Historical Atlas*, Harper and Row, Publishers (Barnes & Noble Books), New York, Copyright © 1964 by Barnes & Noble, Inc.

the force that Rome at first spared for this war was small in numbers and rested entirely upon its own resources, the generals Publius and Gnaeus Scipio, by skillful strategy and diplomacy, not only won over the peoples north of the Ebro and defeated the Carthaginian leader Hasdrubal Barca in his attempts to restore communication with Italy but carried their arms along the east coast into the heart of the enemy's domain.

But eventually their successes were nullified by a rash advance. Deserted by their native contingents and cut off by Carthaginian cavalry, among which the Numidian prince Masinissa rendered conspicuous service, the Roman generals were killed and their troops destroyed (211).

Disturbances in Africa prevented the Punic commanders from exploiting their success. Before long the fall of Capua enabled Rome to transfer troops from Italy to Spain; and in 210 the best Roman general of the day, the young son and namesake of P. Scipio, was placed in command. He signalized his arrival by a bold and successful *coup de main* upon the great arsenal of Carthago Nova (Cartagena) in 209. Though after an engagement at Baecula (Bailen; 208) he was unable to prevent Hasdrubal Barca from marching away to Italy, Scipio profited by his departure to push back the remaining hostile forces the more rapidly. A last effort by the Carthaginians to retrieve their losses with a fresh army was frustrated by a great victory at Ilipa, near Seville, and by the end of the year 206 they had been driven out of Spain.

The Battle of Ilipa

The war in Africa. In 205 Scipio, who had returned to Rome to hold the consulship, proposed to follow up his victories by an attack upon the home territory of Carthage. Though the presence of Hannibal in Italy deterrred Fabius and other senators from sanctioning this policy, Scipio gradually overbore all resistance. He built up a force, which he organized and supplemented in Sicily, and in 204 sailed across to Africa. He was met there by a combined levy of Carthage and King Syphax of Numidia and for a time was penned to the shore near Utica. But in the spring he extricated himself by a surprise attack upon the enemy's camp, which resulted in the total loss of the allied force by sword or fire.

In the campaign of 203 a new Carthaginian force was destroyed by Scipio on the Great Plains 75 miles from Utica, their ally Syphax was captured, and the renegade Masinissa reinstated in the kingdom from which Syphax had recently expelled him. These disasters induced the Carthaginians to sue for peace; but before the very moderate terms that Scipio offered could be definitely accepted, a sudden reversal of opinion caused them to recall Hannibal's army for a final trial of war and to break off negotiations. In 202 Hannibal assumed command of a composite force of citizen and mercenary levies stiffened with a corps of his veteran Italian troops.

After negotiations failed, Scipio and Hannibal met in the Battle of Zama. Scipio's force was somewhat smaller in numbers but well trained throughout and greatly superior in cavalry. His infantry, after evading an attack by the Carthaginian elephants, cut through the first two lines of the enemy but was unable to break the reserve corps of Hannibal's veterans. The battle was ultimately decided by the cavalry of the Romans and their new ally Masinissa, which by a manoeuvre recalling the tactics of Cannae took Hannibal's line in the rear and destroyed it.

The Carthaginians having thus lost their last army again applied for peace and accepted the terms that Scipio offered. They were compelled to cede Spain and the Mediterranean islands still in their hands, to surrender their warships, to pay an indemnity of 10,000 talents within 50 years, and to forfeit their independence in affairs of war and foreign policy.

The Second Punic War, by far the greatest struggle in which either power engaged, had thus ended in the complete triumph of Rome. This triumph is not to be explained in the main by any faultiness in the Carthaginians' method of attack. The history of the First Punic War and that of the Second outside of Italy prove that the Romans were irresistible on neutral or Carthaginian ground. Carthage could only hope to win by invading Italy and using the enemy's home resources against him. The failure of Hannibal's brilliant endeavour to realize these conditions was not due to any strategical mistakes on his part. It was caused by the indomitable strength of will of the Romans, whose character during this period appears at its best, and to the compactness of their Italian confederacy, which no shock of defeat or strain of war could entirely disintegrate.

It is this spectacle of individual genius overborne by corporate and persevering effort that lends to the Second Punic War its peculiar interest. (M.Car./H.H.S.)

The establishment of Roman hegemony in the Hellenistic world. Some scholars believe that Rome made peace with Philip V of Macedonia in 205 in order to concentrate on defeating Carthage but that it intended to settle accounts with him later; others (and this was evidently what Philip himself believed), that Rome intended the settlement to be final. In the latter case, Rome was provoked to declare war against Macedonia (the Second Macedonian War) in 200 by the news brought in 201 by Attalus I of Pergamum (an ally in the first war) and by the Rhodians that Philip had made a secret compact with Antiochus III, king of Syria, to capture and share the external possessions of Egypt, whose new king, Ptolemy V Epiphanes, was a boy.

Provocation for war against Macedonia

Wars against Philip V, Antiochus III, and the Aetolians. Rome's declared object was to force Philip to withdraw Macedonian troops from the three key positions he had garrisoned in Greece—Acrocorinth (since 224), Chalcis in Euboea, and Demetrias on the Gulf of Pagasae. Rome secured the alliance of the Aetolian League; and, with some heart searching because of recent friendly relations with Macedonia, the Achaean League in the Peloponnese; and also of Attalus and Rhodes, whom Philip had provoked in Asia Minor. From the Adriatic, the Romans forced Philip's strongly held position on the Aous River and advanced into Thessaly under Titus Quinctius Flamininus in 198. In 197 they defeated the Macedonian army at the Battle of Cynoscephalae in Thessaly, in which the Roman legion showed its superiority over the Macedonian phalanx.

Peace was made with Philip on generous terms; but the Aetolians, exasperated by being denied the acquisition of towns in Thessaly for which they had hoped, maligned Rome by suggesting that, rather than freeing Greece, the Romans proposed to garrison the country themselves. Flamininus gave the lie to this claim when he proclaimed the liberty of all Greek cities at the Isthmian Games of 196, and the last Roman soldiers left Greece two years later. Rome had kept its word; but some Romans already doubted the wisdom of this move because, with as yet undeclared intentions, Antiochus, who had possessions on the European side of the Hellespont (Lysimacheia, Aenus, and Maroneia) and who had accepted Hannibal as an adviser, was already moving into western Asia Minor.

Negotiations between Antiochus and Rome foundered over the former's refusal to relinquish his possessions in Europe, and the dissatisfied Aetolians solved Antiochus' problem by inviting him, in 192, to save them from the Romans. Antiochus crossed to Greece with inadequate forces and with inadequate supplies to join the Aetolians, who now declared war on Rome. He had no help from Philip, who honoured his peace with Rome. The Achaean League stood by Rome, as did Eumenes II (who had succeeded Attalus as king of Pergamum in 197) and the Rhodians, both of whom expected to profit from Antiochus' defeat.

The Romans, under Acilius Glabrio, defeated Antiochus and the Aetolians at Thermopylae in 191. Antiochus retired to Asia; his navy was soundly defeated in two battles by Roman fleets in cooperation with the Rhodians; and in 190 Lucius Scipio (with his brother, the great Africanus, on his staff) crossed the Hellespont and defeated him at Magnesia. Peace was made, Antiochus being compelled to abandon his possessions in Europe and his territory in Asia west of the Taurus range and retaining only a token navy of 10 ships. The abandoned Aetolians were forced to capitulate on humiliating terms in 189, becoming subject-allies of Rome. That same year the Romans, under Manlius Vulso, conquered the Galatians in Asia Minor.

Rome's policy in the East was fraught with ambiguity. In the Second Macedonian War, the Romans had crusaded

for the "freedom of the Greeks," and, by their own withdrawal from Greece in 194, they had given the appearance of honest crusaders. But although Rome removed Macedonian garrisons from cities in Greece and Asia Minor, Aegina remained the property of Pergamum, having been sold to Attalus in the First Macedonian War, and the Roman peace commission of 197 had even considered giving Eretria in Euboea to Pergamum. After Magnesia, Rome had no compunction about giving a number of Greek cities in Asia Minor to Pergamum and Rhodes.

Establishment of a protectorate over Greece. Considerations of its own interest drove Rome to curtail the aggressive potentiality of Philip, Antiochus, and the Aetolians, with all of whom, after their defeats, Rome was bound by treaty and, so, was justified in acting if terms of these treaties were infringed. Relations with the Achaean League were less clear-cut, and the problem came to a head in 195. The prime object of Achaean policy was to include the entire Peloponnese by the acquisition of Elis, Messene, and in particular Sparta; Sparta's communist regime, established by King Cleomenes III in the previous century, had been revived by Nabis and now evoked envious admiration among the proletariat of the Achaean cities, thus appearing to the governing class in those cities a dangerous contagion that if not stifled might spread. Apprehensive of the slaughter and chaos that might follow an Achaean conquest of Sparta, Flamininus accepted the submission of Sparta in 195 and insisted on its remaining independent; and indeed Rome made a treaty with it.

Relations with the Achaean League

In 194 Philopoemen, a headstrong Achaean patriot who had separated from Flamininus because of mutual jealousy and antagonism, returned after six years from Crete and controlled the league's policy. In 192 he outwitted Flamininus (now back in Greece with a roving commission) and brought Sparta into the league with the promise of no interference with its government. Then, in 188, he settled the issue once and for all with such acts as eliminating the enemies of the league, pulling down the walls of Sparta, abolishing communism, and enslaving the helots once more and putting them up for sale. In 183 Messene seceded from the league in a short-lived revolt, and Philopoemen, seeking to recover it, was captured and killed.

The year 181 was, in the Greek historian Polybius' view, decisive for the future relationship of Rome and the league. Philopoemen was dead, and the Achaean Callicrates persuaded the Senate at Rome that its interest lay in giving uncritical support in the league to the unequivocal supporters, like himself, of the Roman point of view.

Worsening relations with Macedonia

The Roman government carried a considerable responsibility for the worsening of relations with Macedonia and the eventual outbreak of the Third Macedonian War. Philip's refusal to join Antiochus was of untold value, and he was rewarded by the Romans' return of his younger son Demetrius, a hostage in Rome since 197, and the remission of his outstanding indemnity. But Philip felt himself cheated when he was not allowed to keep the Thessalian cities he had taken from the Aetolians when cooperating with Glabrio in 191. Nor was he allowed to recover Aenus and Maroneia on the Thracian coast, which had once belonged to him and were now given up by Antiochus, even though King Eumenes of Pergamum had been granted Lysimacheia and the Thracian Chersonese, which Antiochus was also forced to abandon. Still worse, Rome made an unprincipled attempt to split the Macedonian royal house. Demetrius had won great popularity in Rome when a hostage and, when he returned to Rome in 184 as Philip's representative to answer a variety of charges against the King, the Romans agreed to accept Philip's defense only because Demetrius was his mouthpiece. Neither Philip nor his elder son Perseus could be blamed for thinking that the Romans wished to upset the natural succession. When it was suspected that Demetrius planned to escape from Macedonia to Rome, and when Macedonian envoys returning from Rome in 181 produced a letter from Flamininus that, whether or not it was forged, suggested that Demetrius was conspiring for power, Philip had him killed. By this time Philip had decided on war with Rome; but he died a year later (179), having by his campaigns in the north in the latter part of his reign greatly increased the resources of his kingdom both in wealth and manpower. The Romans believed these campaigns to be exploratory—the prelude to an attempt to invade Italy from the northeast. Philip's son Perseus turned from the north to a policy of reconciliation with the Greeks and, though he was not a soldier of his father's calibre, roused increasing suspicion at Rome, which Eumenes adroitly fostered. Marriage connections suggested that Perseus might become the nucleus of a formidable anti-Roman combination—his own marriage to the daughter of Seleucus IV of Syria and his sister's marriage to Prusias II of Bithynia.

The issue was decided when Eumenes, himself a probable sufferer if Macedonia reemerged as a great power, visited Rome in 172 and persuaded the Senate of Perseus' aggressive plans. The Third Macedonian War was sparked off by a couple of bizarre episodes—an amateurish attempt to kill Eumenes on his way to Delphi while returning from Rome, thought to be instigated by Perseus, and the "evidence" of a prominent citizen of Brundisium that Perseus had invited him to poison eminent Romans who had lodged with him on their way to and from the East. It was a difficult war; the Roman army had to land at Apollonia and cross the mountains into Thessaly. Perseus at the start had the advantage of numbers; indeed he won the first battle, in Thessaly in 171, and after it he offered to treat for peace. The offer was refused; and when, in 170 and 169, the Romans failed to bring Perseus to battle, the belief grew, not only in openly anti-Roman circles, that Perseus might not be defeated after all. The Illyrian chief Genthius joined Perseus. Secret talks were held between Eumenes and Perseus, the full truth of which was never revealed, with the object that, for a suitable price, Eumenes should endeavour to secure a negotiated peace. And the Rhodians, whether or not they had misunderstood Marcius Philippus, the Roman commander in 169, thought that they had been encouraged by him to try their hand at peace negotiations. On June 22, 168 BC, however, the Roman commander Lucius Aemilius Paullus brought Perseus to battle at Pydna and defeated him decisively. Perseus capitulated.

The Third Macedonian War

Perseus lost his throne and survived in exile for two more years at Alba Fucens, near Rome. Macedonia was divided into four independent republics. Genthius had already been defeated in Illyria and was also condemned to exile in Italy. Illyria, like Macedonia, was given its independence; it was divided into three units, but not before 150,000 of its men had been taken and sold into slavery.

The effects of the Roman victory over Perseus were momentous. Eumenes, the Achaean League, and Rhodes were all suspect. Although Attalus II, brother of King Eumenes, was received in Rome in 167, Eumenes himself was met by a quaestor at Brundisium and told that the Senate would receive no kings in Rome—a snub that was reinforced by the audience already given to his enemy, King Prusias of Bithynia, who in Rome startled the Senate by doing obeisance to them and addressing them as saviour gods. The Achaeans were required to send 1,000 hostages to Rome; they were relegated to country towns in Italy until 151, when the 300 survivors were allowed to go home. One of the hostages, Polybius, did well, living as a favoured guest of the Scipios and laying the foundations of his great history.

Rhodes fared worst of all. Already in 177 it had discovered the ambivalence of Roman policy: having been "given" Lycia and Caria on the mainland by the Romans after Magnesia and having proceeded to their direct administration, it was informed by Rome, in consequence of a Lycian appeal, that it had not been given Lycia and Caria at all but had simply been bound to them in friendship and alliance. Now, in 168, Rhodes's delegates, sent to propose a peaceful settlement of the issue with Perseus, arrived in Rome to learn that the war was over. A Roman praetor in 167 proposed that Rome should declare war on Rhodes, but the censor Cato pointed out that whatever the Rhodians' sympathies in the recent war might have been, they had performed no act of overt hostility. So the Rhodians received and forcibly obeyed Rome's harsh

commands—to abandon Caunus and Stratoniceia on the mainland, neither of which had been received from Rome, and to withdraw from Caria and Lycia. Delos was declared a free port, and Rhodes consequently lost about six-sevenths of its income from harbour dues. Then, having reduced Rhodes to a second-class power, Rome consented, in 165, to make a formal alliance with her.

Piracy, based largely on Crete, was always a menace, and Rhodes, with Syria, had been the policeman of the eastern Mediterranean. With the Syrian navy already reduced to a token force and Rhodes now weakened, piracy flourished unchecked; and in the following century the Romans were themselves the victims of their own shortsighted policy.

Establishment of provinces in Africa and the East. Further east, as part of their policy of preserving a balance of power between Syria and Egypt, the Romans in 168 stopped the invasion of Egypt by the Seleucid ruler Antiochus Epiphanes. A Roman legate, Gaius Popillius Laenas, caught up with him and ordered him home, and the King did as he was told. Between 151 and 146 Rome found itself at war with Carthage, Macedonia, and the Achaean League, as a result of which all three powers lost their independence; and the provincial empire of Rome, previously confined to the west, was extended into Africa and into the Balkans.

In Africa, Carthage studiously respected the peace treaty of 201 despite constant filching of territory by its Numidian neighbours under King Masinissa and by the hostile decisions of successive Roman boundary commissions. Its commerce and material resources revived in the 2nd century with such rapidity as to excite the jealousy of the growing mercantile population of Rome and the alarm of its more timid statesmen. Under the influence of these feelings the conviction—sedulously fostered by Cato the Censor—that "Carthage must be destroyed" overbore the scruples of more clear-sighted statesmen. A *casus belli* was readily found in a formal breach of the treaty, committed by the Carthaginians in 150 when they resisted Masinissa's aggressions by force of arms. A Roman army was dispatched to Africa; and although the Carthaginians consented to make reparation by giving hostages and surrendering their arms, they were goaded into revolt by the further stipulation that they must emigrate to some inland site where they would be debarred from commerce.

Cato's influence

By a desperate effort they created new war equipment and prepared their city for a siege (149). The Roman attack for two years completely miscarried until in 147 the command was given to a young officer who had distinguished himself in the early operations of the war—Scipio Aemilianus, the adopted grandson of the former conqueror of Carthage. Scipio made the blockade stringent by walling off the isthmus on which the town lay and by cutting off its sources of supplies from overseas. His main attack was delivered on the harbour side, where he effected an entrance in the face of a determined and ingenious resistance. The struggle did not cease until he had captured house by house the streets that led up to the citadel.

Final destruction of Carthage

Of a city population which, according to Strabo, had exceeded a quarter of a million, only 50,000 remained at the final surrender. The survivors were sold into slavery; the city was razed to the ground and its site condemned by solemn imprecations to lie desolate forever. The territory of Carthage, which had recently been much narrowed by Masinissa's encroachments, was converted into a Roman province under the name of "Africa."

Macedonia too, with Illyria and Epirus and with Achaea as an appendage, became a Roman province in 146. The Macedonians had missed their monarchy, and when, in 149, a certain Andriscus claimed that he was the son of Perseus, he found followers and was soon in possession of the whole country. Andriscus was defeated in 148 by Quintus Caecilius Metellus (who won the name "Macedonicus") and was brought to Rome and executed.

Appropriately enough, given past history, the independence of Sparta was the issue that brought an end to the Achaean League. Callicrates had died, and control of league policy passed to fanatical opponents of Rome—Diaeus and Critolaus. Sparta had seceded, and in 147 a Roman commission announced its decision that Corinth and Argos should be detached from the league as well. Negotiation was impossible in the consequent atmosphere of insensate fury. The league was ready for war. The Roman leader Metellus Macedonicus defeated Critolaus north of the isthmus; Lucius Mummius, consul of 146, arrived and defeated Diaeus. Corinth was pillaged, its wealth of artistic treasures plundered and dispatched to Rome. The city was destroyed and abandoned.

The independent kingdom of Pergamum survived until 133, when its last king, Attalus III, died and, having no heir, bequeathed his territory to Rome. With the title of "Asia," it became Rome's seventh province.

Beginnings of Roman provincial government. The Roman Empire was founded and developed as Rome successively assumed direct responsibility for the government and administration of conquered territories (provinces): of Sicily in 241, of Sardinia and Corsica (always administered as a single province) in 238, of Spain (the two provinces of Nearer and Farther Spain) in 197. It was at the start a western empire. Then, after half a century, came Macedonia (with Illyria and Achaea) and Africa in 146, and finally Pergamum (the province of Asia) in 133. Each province had a separate written constitution (a *lex provinciae*) with detailed regulations made by the conquering general in association with a commission of ten senators (*decem legati*) acting within broad regulations laid down by the Senate. This constitution named the cities within the province that retained their independence, uncontrolled by the governor's jurisdiction (*civitates liberae*); they were usually cities that had assisted Rome at the time of the conquest and that in some cases were already bound to Rome by treaty (*civitates foederatae*). The constitution also specified the nature of the taxation, whether direct (*stipendium*) or indirect (tithes [*decumae*] on agricultural production, which were the rule in Sicily and Sardinia).

Development of an empire

At the head of the Roman administration were the annually appointed governor and the financial secretary (one of the quaestors of the year). The governor's function was expected to be predominantly administrative (he would constitute the province's high court of justice) and comparable to that of a praetor in Rome. Because the governor was normally a praetor, the number of praetors was, therefore, increased to four in 227 and to six in 197, lots being cast each year to determine the postings. But because the Hannibalic war and the great wars in the East made other demands on praetors (*e.g.*, naval command) it was frequently necessary to extend a governor's provincial command for more than a year. His power (*imperium*) remained the same, but after his magisterial year he was a propraetor or, more properly, a proconsul.

Only when there was a major war in a province, or the prospect of one, was a consul sent out to be its governor. This happened ten times in the case of Nearer Spain and four times in Farther Spain in the period 153–134 BC, and it happened in Sicily after the outbreak of a serious slave revolt in 134 and 133.

Consciousness of imperial responsibility, which rested with the Senate, grew slowly and was, perhaps, never fully grasped (except in rare cases, such as by the elder Cato and the elder Tiberius Gracchus) until, under the empire of the Caesars, the responsibility was assumed by the emperor; and only when faced with undeniable evidence of misgovernment and corruption did the Senate take corrective measures. In 149, after a scandalous acquittal in a public trial of Servius Sulpicius Galba, a criminally corrupt governor of Farther Spain, a permanent court (the *quaestio repetundarum*) was established in Rome, with senators as jurors, before which provincial subjects—provided that a Roman would act as their patron—could impeach Roman officials for extortionate practices.

The Roman economy was quick to feel the benefits, however unethical, of the great wars and of the new provinces. Apart from the wealth brought back to Rome in official and private loot, there were the indemnities from Carthage, Philip, and Antiochus. The grain paid in taxation in Sicily and Sardinia was sold in Rome at a price with which homegrown grain could not compete, with disastrous effects on large sections of Roman agriculture. Rome also gained wealth from the Spanish mines.

In 167 the property tax for Roman citizens in Italy was abolished—evidence of imperialism's dividends.

Apart from brigandage, which had always been endemic in Sardinia, the island provinces, until the great Sicilian revolt of 134, caused little anxiety to the Roman government. Spain, however, was a different matter. Fighting in the mountains was difficult, the policy of resettling communities in cities in the plains was hard to enforce, and the peninsula was never completely conquered and pacified until the time of Augustus. From 154 there was continuous war against both the Lusitani in Farther Spain and against the Celtiberi in Nearer Spain—a war that produced a Spanish resistance leader of genius, Viriathus. After his murder in 139, the war continued until at last Scipio Aemilianus captured the stronghold of Numantia in 133. The war had its repercussions at home. Military service in Spain was exceedingly unpopular, and in both 151 and 138 tribunes imprisoned the consuls, so as to prevent the levies (behaviour that was to point the way to the Gracchi).

Establishment of colonial outposts

Of the 70 consuls between 200 and 166 BC, 55 were engaged for part of their year of office in campaigning with armies in the north of Italy against the Gauls, particularly against the Ligurians in the northwest and the Boii in the northeast, with the object of clearing the northern Apennines and bringing the peninsula up to the Alps under firm control. Once a district was sufficiently pacified, Latin colonies were founded at strategic points, manned by Latins (and also, sometimes, by Romans) who were attracted by the prospect of free land even in potentially dangerous areas. And great roads were built—always a vital element in Romanization—for communication and, in particular, for the rapid movement of troops.

Particularly in the northeast, fighting was sometimes critical, sometimes a mere matter of triumph hunting (a "Ligurian Triumph" became an expression of mockery), and the serious fighting ended after Marcus Baebius Tamphilus, consul in 181, removed 40,000 Apuani with their families to the neighbourhood of Beneventum in southern Italy. A Latin colony was settled at Luca in 180, and a Roman colony at the port of Luna in 177. In the northeast a Latin colony was sent to Bononia (Bologna) in 189, and another, with unusually large allotments of land, to Aquileia in 181.

The names of the two consuls of 187 are perpetuated in the Via Flaminia from Arretium (Arezzo) to Bononia and in its continuation, the Via Aemilia, from Bononia to Placentia. In 175 Bononia was connected by road with Aquileia, and in 148 a road was driven diagonally across the peninsula from the Mediterranean through the Po Valley to the head of the Adriatic—the Via Postumia, which ran from Genna (Genoa) through Placentia, Cremona, and Verona to Aquileia.

In southern Italy a number of small citizen colonies, each of 300 settlers (coast-guard colonies), were dispatched to vital points on the coast after the Hannibalic war to strengthen coastal defense, particularly in view of fears of a possible invasion of Italy by Antiochus' fleet under the command of Hannibal. These were founded in 194: Puteoli, Salernum, Volturnum, Liternum, Sipontum, Buxentum, Croton, and Tempsa.

Roman government, economy, and culture. Polybius, deeply experienced in the political life of the Hellenistic world, completed, during and after the time he was a hostage in Rome, a vast history of the years 264 to 146 BC that shows how Carthage and Macedonia during this period were absorbed into Rome's empire and how the rest of the Hellenistic world fell under the dominating influence of Rome. He tried to determine the secret of Rome's uninterrupted success. He might, as did Philip V (in a letter known from an inscription), have emphasized the liberality with which, in contrast to Greek city-states, Rome increased the number of its citizens through granting citizenship to slaves and aliens, and the manner in which it often resettled, instead of enslaving, conquered peoples. Or he could have stressed the system of *clientela* (the dependence of the poor and unprivileged on a rich and distinguished patron, which was reflected in public life when they recorded their votes), including foreign *clientelae* (the protection given to the interests of provincial subjects and foreign kings by particular Roman families of prominence). But he did not. He emphasized the binding force of Roman public religion as inducing in Roman public servants, through the sacrosanctity of oaths, an integrity that had no parallel in contemporary Greece. He was impressed by Roman loyalty to the state and, within the great families, loyalty to their own distinguished traditions. He stressed the unrivalled efficiency of the Roman army as a fighting force.

Elements of government. Above all, Polybius was struck by the manner in which the three elements—of monarchy (the consuls), aristocracy (the Senate), and democracy (the popular assemblies)—struck a balance of power that guaranteed a stable equilibrium; this is not an entirely attractive theory, however, because it greatly exaggerates the independent power and importance of the two annual consuls, whose position was anything but monarchical.

Qualifications for public service

The Roman sovereign body was the Senatus Populusque Romanus (Senate and the Roman People). The Senate's membership depended on the two censors—very senior senators elected normally at five-year intervals. They admitted to the Senate those who had held the quaestorship within the preceding five years (*quinquennium*), and then they went through the list of existing senators, expelling those who had acquired any public notoriety (*infamia*), as Cato expelled the brother of Titus Quinctius Flamininus in 184 for scandalous behavior when campaigning in northern Italy. The revised list of senators was then published, the first name on the list being that of the *princeps Senatus,* who was invited to speak first in any debate at which he was present. Publius Scipio Africanus was *princeps Senatus* from 199 until his death, in 183; Marcus Aemilius Lepidus, from 179 until his death, in 152.

In the early 2nd century rules were made governing the succession of offices in a public career (*cursus honorum*). The first such rule stated that no one could be elected consul if he had not already been praetor. In 180 minimum ages were established for holding office—for plebeians 36 was the minimum for the curule aedileship, 39 for the praetorship, and 42 for the consulship; patricians, it seems, could hold the magistracies two years earlier. The quaestorship could be held after the 10 years' cavalry service done by youths of the upper classes (as against 16 years' infantry service to which others were liable). This service started at age 17; so the quaestorship could be held at age 27, though there are recorded cases when it was held at 25. Whether the quaestorship was an essential part of the *cursus* is uncertain. Tenure in the curule aedileship and the tribunate of the plebs were not compulsory; but the ambitious rarely omitted them because they offered an opportunity to win popularity, which was likely to be reflected in the voting when candidates stood for the praetorship. (The curule aediles were responsible for the public games, and the tribunes of the plebs had the opportunity to champion popular discontents; for instance, during the calling-up for the wars in Spain in 151 and 138, the tribunes stopped the enrollment of soldiers by imprisoning the consuls.) Beginning in 267 there were eight quaestors, whose duties were largely concerned with public finance; there were four aediles (two plebeian, two curule), and 10 tribunes. Before 227 there were two praetors (the second, *praetor peregrinus,* since 242), who were wholly occupied with judicial duties in Rome; after 227 there were four, and after 197 there were six, the additional praetors being required as provincial governors.

There were two consuls; therefore, only one in four of those elected to the quaestorship could eventually reach the consulship and so ennoble his family, if it was not noble already (the *nobiles* were families that could number a consul among their ancestors).

If both consuls were in Rome, they held the *fasces* (badge of authority) and presided in the Senate on alternate months, the consul who had headed the poll presiding in the first month of the year (March until 153 and January after that). In the period of the great wars, however, the consuls generally left Rome early in the year, after they had discharged their necessary formal obligations (for instance, the celebration of the Latin games on Mons Al-

banus), in order to command armies on campaign; they normally returned at the end of the campaigning season. In the interval, the urban praetor presided over meetings of the Senate.

A consul was responsible for the elections, which took place late in the year. If the consul was unable to return to Rome for this purpose, he could nominate a dictator—a man who held office for three *nundinae* (probably 17 days) and had no other function to perform. If both consuls died before the elections or if the year ended without the elections taking place, there was an *interregnum* (sovereignty reverting to the patrician members of the Senate, as had happened in early days after the death of a king). An interrex then held the election.

A consul or praetor, on the termination of his magistracy, could retain his *imperium* with the title of promagistrate (proconsul or propraetor); such extension of *imperium* and use of promagistrates developed quickly during the Second Punic War because the number of necessary military commands exceeded the number of available magistrates. The former consuls Publius Cornelius Scipio and his brother, for instance, commanded in Spain as proconsuls until they were killed in 211. This practice was continued and extended to meet the exigencies of provincial government in the second and first centuries.

Political debates in the Senate

Although there were public meetings (*contiones*), it was in the Senate that political debates on the highest issues of state took place. Members were invited to speak in order of seniority, a fact that explains the great weight and importance of the former consuls (the *consulares*) in the Senate. In the middle republic these were members of a limited number of the great families whose influence, exerted through their *clientes,* assured their election—consuls and praetors were elected by the Comitia Centuriata, in which the rich exercised a predominant influence. The results of the elections show clearly that the presiding magistrate was able to exert great influence in favour of candidates who were related to him or who were his political associates.

In the Senate the consulares did not speak with a single voice, and there was strong infighting, certain families—for example, the Cornelii Scipiones and the Aemilii, and the Claudii and the Fulvii—being closely linked in political friendship (*amicitia*) with one another and in opposition (*inimicitiae*) to other groups. Such groupings, however, were not rigid, for groups dissolved and reformed around particular prominent political figures and the policies favoured by those politicians. The Scipios, for example, favoured considerate and friendly relations with the Hellenistic kings and welcomed the genial influence of Greek civilization on Roman life, which Cato opposed. In the years around 170 a new ruthlessness marked Roman foreign policy, and this seems to have been largely due to the influence of the Fulvian group, including Marcius Philippus and Popillius Laenas, who were then in a controlling position in Roman politics.

Some of the personal antagonism of leading members of the great families was shown in the opposition of Fabius Maximus to Scipio Africanus, and it assumed great virulence early in the 2nd century. There was an attempt to deny Acilius Glabrio, consul of 191 (a new man and a protégé of the Scipios), a triumph for his great success in Greece, and soon afterward there came an attack on the integrity of Africanus and his brother Lucius in their dealings with Antiochus. In consequence, both Acilius and Lucius failed to be elected to the censorship, and Africanus spent his last years at Liternum in Campania in a kind of banishment. Similar attacks were launched against Manlius Vulso after his campaign in Asia Minor in 189. This kind of personal hostility could harm Roman interests if a commander in the field was prepared to make peace with an enemy on disadvantageous terms rather than allow his successor the credit of victory, or if a commander tried to overturn a peace so as to continue a war under his own command: in 201 the consul Gnaeus Cornelius Lentulus tried, without success, to overturn the peace with Carthage made after Zama.

Thanks to its cool-headed direction of the Second Punic War, the power of the Senate increased insensibly. In the alarm caused by the exposure of the Bacchanalian conspiracy in 186, nobody objected to the Senate's behaviour in empowering the consuls to put even citizens to death, though by law a citizen could only be executed after trial and condemnation by the People. Similar behaviour by the Senate in the case of Tiberius Gracchus' supporters in 132 was to be exposed as unconstitutional by Gaius Gracchus nine years later.

The Senate's role in war

The declaration of war and the making of peace were still the formal responsibility of the People, as was shown particularly in the declaration of the First Punic and Second Macedonian wars; but the administrative arrangements that followed peace were in the Senate's hands. If the Senate decided that a conquered land should be made a Roman province, the Senate itself decided the general lines on which the new province should be organized and, as has been seen, sent out ten of its members to collaborate with the victorious general on the spot in giving detailed effect to its general resolution, as well as to issue the charter of the province.

The oldest of Rome's effective assemblies, the Comitia Centuriata, was not a democratic body. Of the 193 voting units (after the reform of 241), 70 belonged to the highest of the five property classes and 18 to the aristocratic cavalry (*equites equo publico*), so that the rich came near to having an absolute majority of votes. A junior century of the top property class (*centuria praerogativa*) voted first; and its vote, when announced, influenced the voting of the other centuries. The legislative Comitia Tributa, which also elected the quaestors, was more democratic because the voting unit was the tribe; but in the 3rd century BC few members of the 31 country tribes, apart from the rich, could spare the time to come to Rome to vote, so the result was the same. During the 2nd century, however, a big change took place; the population of Rome grew rapidly on account of the drift to the city of the ruined peasantry, who had lost their farms through the spread of large grazing estates (*latifundia*). These immigrants remained members of their original country tribes, so that by the middle of the 2nd century the Comitia Tributa (and likewise the Concilium Plebis) became representative bodies of the city dwellers of Rome itself.

In the matter of legislation, the Comitia Centuriata and the Comitia Tributa could vote only on measures that had already been approved by the Senate. The Concilium Plebis, on the other hand, under the presidency of the tribunes, had received in 287 the right to legislate for the whole Roman people without consultation of the Senate; and the way seemed open to extreme democracy. This development was arrested by the invasion of Pyrrhus (280) and the succession of great wars. A land bill of Gaius Flaminius in 232, carried against the wishes of the Senate, was a rarity. Normally, if conservative interests were threatened by a revolutionary, or even a reforming, tribune, it was not difficult to persuade one of his nine colleagues, with promises of support in his political future, to veto a proposal. This was the simple means by which Tiberius Gracchus' opponents expected to kill his land bill in 133, and the bill would have failed if Tiberius Gracchus had not been prepared to act in violation of the Roman constitution by persuading the plebs to unseat the dissenting Octavius.

As long as there was an open system of voting, without ballots, the rich and powerful, acting through their agents, were able by intimidation to control the voting of their clients. There was, therefore, strong conservative opposition to the introduction of voting by ballot, with its attendant secrecy, for elections in 139 (Lex Gabinia Tabellaria) and for legislation in 131 (Lex Papiria Tabellaria).

By very early Roman law, constantly reenacted, no Roman citizen could be executed except after condemnation by the People; and so the People in the Comitia constituted the supreme criminal court of Rome. Impeachment could be by a tribune or an aedile; this popular function was supplemented, and in the end replaced, by the creation of permanent criminal courts, sitting under a praetor, on the model of the *quaestio repetundarum,* the permanent court to handle extortion cases, established by a bill of the tribune Calpurnius Piso in 19.

There was another field in which the People were not altogether voiceless or ineffective. It was at a meeting of the Comitia Tributa under the presidency of tribunes that, despite his youth and inexperience, young Publius Cornelius Scipio was sent to command in Spain in 210; and it was as a result of popular outcry in favour of disregarding legal requirements concerning the minimum age for the consulship that even the Comitia Centuriata in 148 decided that Scipio Aemilianus, a candidate for the aedileship, should instead be elected consul and sent to conduct the war against Carthage. Thus there were 3rd- and 2nd-century precedents for the extraordinary popular commands of the last decades of the republic.

Economy. By the Lex Claudia of 218, senators and their sons were effectively precluded from trading by being forbidden to own vessels larger than those necessary for shipping produce from their estates to Rome; but it would be a mistake to think of them as utterly distinct from the business and trading class, for they might have relatives who were in business, and even the senators themselves might carry on enterprises in the name of one of their freedmen. Nonetheless, the Senate seemed to show a complete lack of concern with imperial commercial expansion, as evidenced by the closure of the Macedonian mines in 167 and the destruction 21 years later of two of the finest harbours in the Mediterranean—Corinth and Carthage.

Private enterprise

At the same time there were large business interests operated by a class of men that later developed into the equestrian order. As has been seen, during this period many of the great roads of Italy were constructed; and, indeed, the first of the great imperial roads—the Via Egnatia from Dyrrhachium and Apollonia to Thessalonica—was built not long after 146. Tenders for such work were submitted to the censors who, after scrutiny, allotted the contracts. In Spain the state soon abandoned its first attempt to manage the mines and handed the management over to private enterprise, the contractors paying a rent to the state. There were contracts for the arming and equipping of the Roman armies, which must have required particularly large quantities of leather. After Cannae the companies (*societates*) of contractors agreed to defer payment on condition of being fully insured against losses by storm or war.

The greatest economic change in Italy was the rapid development (in the early 2nd century) of ranching as a major industry, particularly in Campania and south Italy, which had been devastated and depopulated as a result of Hannibal's occupation. Rich men extended their holdings by encroaching on public land (*ager publicus*), partly by buying up at a low price the small farms of those who, largely through extended periods of military service (particularly in Spain), were no longer able to work their holdings at a profit. Thus, large estates (*saltus, pascua*) came into existence, partly in hill country for summer grazing, partly in the plains for winter grazing. In the south, sheep were raised, farther north, oxen, the hides of which produced the leather that was so badly needed.

These enterprises were assisted by the ready availability of cheap slaves, captives from the wars in the East. The slaves were tough men whose labour was ruthlessly exploited; they were often chained, and they slept in slave prisons (*ergastula*). Large numbers were required—to fight off wild beasts and to prevent cattle rustling.

These great grazing estates, the *latifundia* that, in the elder Pliny's famous expression, "were the destruction of Italy," had nothing in common with the medium-sized and large arable farms, on the economy and management of which the elder Cato wrote his *De agricultura.* Cato, indeed, discussing means of making a fortune, dismissed commerce as risky and moneylending as being disreputable in itself and said that more money could be made by ownership of even a poor ranch than by agriculture.

Demographically, the swelling of the population of the big towns, particularly Rome, by immigration of dispossessed smalll farmers was a striking feature of the first half of the 2nd century and was the movement which Tiberius Gracchus later (133) tried to reverse. In the country towns the immigrants made what money they could as casual labourers, and it is doubtful that there was much unemployment among the expanding population of Rome. Grain production in the vicinity of Rome and the large coastal towns of Italy dropped heavily as a result of the import of tax grain from Sicily and Sardinia; but in the north agriculture flourished in the rich plain of the Po Valley (Gallia Cisalpina).

The rapid growth of luxury imports and also of the number of slaves from the eastern Mediterranean brought Roman and Italian businessmen in large numbers to the East, particularly to Delos (made a free port in 166), where a large Roman and Italian business community came into existence; many thousands of persons a day are said to have passed through the Delos slave market. There were numbers of Roman settlers in the western provinces, too. Wounded Roman soldiers were left behind at Italica in Spain by Scipio Africanus in the Second Punic War. In 171 about 6,000 sons of Roman soldiers who had married Spanish women established the first Latin colony outside Italy, at Carteia, near Gibraltar.

Culture. From the 5th century onward Rome had received mention by Greek historians; after its defeat of Pyrrhus, when Ptolemy II made diplomatic overtures to the city, its formidable potential was evidently recognized in the Hellenistic world, and soon after that Timaeus wrote at length of Italy and Rome in his Sicilian history. By 217 mainland Greece was frighteningly aware of "the cloud in the West"—the victor, whichever it should be, of the Second Punic War. Both of the belligerents were "barbarians" in the technical Greek sense of the word; Rome's very name was also the word in Greek for "brute force" (rhōmē).

Hellenizing influences on Rome

In sharp contrast was the impact the Greek world made on the Romans, first in Campania in the 4th century, then in Sicily and Greece. There were magnificent temples, centuries old, permanent stone theatres, public baths (hot baths at that), and luxurious private houses. There were performances of classical and modern tragedy and comedy in the theatres. There were exercises and athletics in the palaestrae of the baths and at public festivals, with young men competing naked—"Greek games," by which Romans were always a little shocked. There was the Greek language, infinitely more delicate and flexible than contemporary Latin; there was the treasure-house of Greek literature, from Homer onward; and there was the deeply embedded Greek interest in philosophy, in argument, in questioning the truth of what, to the Romans, were fundamentals—traditions of religion and government. Rome had to come to terms with this newly discovered dynamic and to determine if it was to enrich Roman life or to serve as a catalyst.

The children of upper class Roman families now learned to speak Greek—often from a Greek slave nurse—at a very early age; and Greek literature was taught, either in school or by a family tutor, a Greek freedman or slave. As a result, the educated Roman of the 2nd century BC was bilingual and was fluent in both languages. It has been suggested that in the family of the younger Scipio, Greek, not Latin, was the language of general conversation.

Greek literature offered a challenge that was accepted. The Roman poet and dramatist Lucius Livius Andronicus (*c.* 284–*c.* 204) published a Latin version of Homer's *Odyssey;* Gnaeus Naevius (*c.* 270–*c.* 199) wrote a Latin epic on the First Punic War, set against Rome's early background; and Quintus Ennius (239–169), a Calabrian, wrote an epic history of Rome to 171 BC—the *Annales.* Roman drama made its debut as an adaptation of Greek, though it sought to be distinctively Roman. Livius Andronicus was the first playwright. Naevius, Marcus Pacuvius (*c.* 220–*c.* 130), who came from Brundisium, and Pacuvius' uncle Ennius were prolific authors of tragedies. Ennius and Gaius Lucilius (*c.* 180–103/102), respectively, gave a start to the distinctively Roman genres of affable and of hard-biting satire. There were comedies, too—Roman adaptations of Greek Middle Comedy but with a distinctively Roman veneer—in this genre Plautus (born *c.* 254), an Umbrian, and Terence (*c.* 186/185–159), who was of Libyan stock, were preeminent (21 comedies of Plautus survive, six of Terence). Plays were acted at Roman festivals on improvised platforms in front of temples, with temporary seats erected for the audience. (There was no permanent stone theatre at Rome until 55 BC.)

While, with the exception of Lucilius, Rome's earliest literary figures were men of obscure origins, the first writing of history by Romans was done by politicians (the earliest of them, Quintus Fabius Pictor, flourished at the end of the 3rd century). And so a lasting Roman tradition was established that historical writing was a field for men with experience in public affairs. Based on briefly annotated lists of magistrates preserved by the pontiffs and on the vainglorious traditions of the great families, early Roman history before the invasion of Pyrrhus was largely fictional. The first histories were written in Greek, partly to counter hostile accounts of Rome by contemporary Greeks, including accounts of the Punic Wars written from the Carthaginian standpoint. Cato's history (*Origines*), however, was written in Latin and set the fashion for the future.

Greek religion both startled and frightened the Romans; unlike Roman religion, it was anthropomorphic, and the Greek world was full of statues—often of the highest art—of Greek gods and goddesses in the image of men and women. At the same time, contemporary Greek skepticism shocked the fundamental piety of the Romans. On the instruction of the Oracle at Delphi, a black stone from Pessinus in Asia Minor was sent to Rome in 205–204 and, worshipped as the Great Mother, became an object of Roman cult, though with Oriental, not Roman, priests. New orgiastic cults seeped in; and in 186 there was panic on the discovery of the existence of a widely diffused cult of Bacchus, which seemingly involved not only unspeakable immorality but also a plot to overthrow the established government. A witch hunt and a spate of executions followed before the cult was transformed to one of closely supervised respectability, not only in Rome but throughout Italy.

Dangers to the traditional life

The firm stability of Roman traditional life was endangered, conservatives thought, not only by such orgiastic cults and by the increasing popularity among the wealthy of luxurious homes (with hot baths) but also by the acute logic of Greek philosophers who used reason to destroy belief and by the persuasive arts of Greek rhetoricians. In Rome they shocked the old, but they fascinated the young. Thus, when three distinguished Greek philosophers arrived in Rome in 155 to urge the return of the surviving Achaean hostages and conducted widely attended public discussions, Cato had them bundled out of the city. Greek ideas of social equality were utterly un-Roman—ideas by which, it would be claimed, Tiberius Gracchus was corrupted. On the other hand, Stoic ethics were supremely adapted to suit the best Roman mores.

Two schools of thought existed from the start. Both Scipios were patrons of the new writers—Scipio Aemilianus had such Greeks as Polybius and the Rhodian philosopher and historian Panaetius as his friends; Cato opposed them, himself tutoring his son rather than employing a Greek tutor. But Cato was not a fool; he learned Greek as did everybody else. Yet, morbidly fascinated by the spectre of their own decline, later Romans accepted the commonplace that Eastern conquest had weakened the tough Roman fibre, in particular because it introduced new and extravagant tastes.

Only in law did Greece exercise little influence on Rome. Based on the Twelve Tables that schoolboys had to learn by heart, Roman law was amplified by successive laws and senatorial resolutions (*jus civile*), by the interpretations of successive praetors (*jus honorarium*), and by the teaching of the jurists—the earliest of them Tiberius Coruncanius, consul of 280 BC and the first plebeian *pontifex maximus* ("chief priest"). (J.P.V.D.B.)

THE LATE REPUBLIC (133–31 BC)

The fall of Carthage and Corinth did not mark even a temporary end to warfare. For a few years the interminable wars in Spain dragged on—against the Lusitanian leader Viriathus (who had rallied that tribe and won some important victories) in the south and against the Celtiberi based on Numantia in the centre of the peninsula. The sorry tale of Roman defeats, treaties concluded and broken, perfidy, and genocide continued to mark these wars as no others in republican history. But Roman methods finally triumphed. Viriathus was assassinated, and resistance in the south soon ended. The Numantines forced a Roman consular army under Gaius Hostilius Mancinus to surrender in order to escape destruction, and Mancinus promised them peace (137 BC)—an act that inevitably was disavowed by the Senate. Finally, in 134 Publius Cornelius Scipio Aemilianus—again elected consul—enrolled an army consisting chiefly of private clients from all over the Roman world (thus avoiding the internal troubles caused by major levies for the unpopular Spanish front) and finished the war with the capture and destruction of Numantia, whose inhabitants were killed or enslaved (133). The long wars were over at last, and although the picture of peaceful development in Spain must not be exaggerated (five Spanish triumphs are recorded in the 50 years between the time of Scipio and that of Sulla), Spanish wars were no longer a running sore in Roman political life.

End of the wars in Spain

The aftermath of the victories. War and military glory were an essential part of the Roman aristocratic ethos and, hence, of Roman political life. Apart from major wars still to come, small wars on the frontiers of Roman power—never precisely fixed—continue to provide an essential motive in Roman history: in Spain, Sardinia, Illyria, and Macedonia, barbarians could be defeated and triumphs won. Thus the limits of Roman power were gradually extended, and the territories within them pacified, while men of noble stock rivalled the *virtus* (courage and qualities of leadership) of their ancestors, and new men staked their own competing claims, winning glory essential to political advancement and sharing the booty with their officers and soldiers. Cicero could still depict it as a major disgrace for Lucius Piso (consul 58 BC) that he had won no triumph in the traditionally "triumphal" province of Macedonia. Nonetheless, the coincidence of the capture of Corinth and Carthage was even in antiquity regarded as a turning point in Roman history: it was the end (for the time being) of warfare against civilized powers, in which the danger was felt to be greater and the glory and the booty were out of proportion to those won against barbarian tribes.

Provincial administration. The first immediate effect was on the administration of the empire. The military basis of provincial administration remained: the governor (as he is called) was in Roman eyes a commander with absolute and unappealable powers (*imperium*) over all except Roman citizens, within the limits (his *provincia*) assigned to him (normally) by the Senate. He was always prepared—and in some provinces expected—to fight and win. But it had been found that those unlimited powers were often abused and that Senate control could not easily be asserted at increasing distances from Rome. For political and perhaps for moral reasons, excessive abuse without hope of a remedy could not be permitted. Hence, when the decision to annex Carthage and Macedonia had been taken in principle (149 BC), a permanent court (the *quaestio repetundarum*) was established at Rome to hear complaints against former commanders and, where necessary, to assure repayment of illegal exactions. No penalty for offenders was provided, and there was no derogation from the commander's powers during his tenure; nevertheless, the step was a landmark in the recognition of imperial responsibility, and it was also to have important effects on Roman politics. Another result of the new conquests was a major administrative departure. When Africa and Macedonia became *provinciae* to be regularly assigned to commanders, it was decided to break with precedent by not increasing the number of senior magistrates (praetors). Instead, prorogation—the device of leaving a magistrate in office *pro magistratu* ("in place of a magistrate") after his term had expired, which had hitherto been freely used when emergencies led to shortages of regular commanders—was established as part of the administrative system: thenceforth, every year at least two praetors would have to be retained as promagistrates. This was the beginning of that dissociation between urban magistracy and foreign command that was to become a cardinal principle of the system of Sulla and of the developed Roman Empire.

Social and economic ills. It is not clear to what extent the temporary end of the age of major wars helped to produce the crisis of the Roman Republic. The general view of thinking Romans was that the relaxation of external

pressures led to internal disintegration. This has happened in other states and the view is not to be lightly dismissed.

Intensified poverty and discontent

Moreover, the end of large-scale booty led to economic recession in Rome, thus intensifying poverty and discontent. But the underlying crisis had been building up over a long period. In view of the inadequacy of the source material and the still limited results of archaeological research, its nature cannot be confidently defined. But some facts are certain. Large tracts of southern Italy had been devastated in the Hannibalic war and, partly for ecological reasons, never recovered. Little successful resettlement took place on the land confiscated there by Rome from rebellious allies after victory, and much of it was turned over to grazing, chiefly by transhumance methods, hence chiefly exploited by wealthy Romans. Nearer Rome, imports of provincial wheat destroyed local wheat growing, with the result that peasants were driven off the land and flocked into the city, while their holdings were bought up by wealthy neighbours and often consolidated into medium-sized mixed farms. It was for this class of men that Cato wrote his book *De agricultura.* Farther north, causes and effects alike were less catastrophic; but long wars overseas and rigorous conscription led to the decline of peasant holdings and the accumulation of land. A population drift to the cities (both Rome and the country towns) is noticeable by mid-century, with the towns apparently growing and prospering and, no doubt, attracting peasants, who, by selling out and moving there, would cease to be landowners and liable to the draft. Their divorce from the land, however, was not absolute: seasonal work in agriculture always helped to provide a livelihood for former peasants and their families.

Rich landowners probably put tenants in the place of the peasants they had bought out; if progressive, they might change over to Cato's kind of mixed farming, based on slave labour and supplemented by seasonal free labour as required. Rich men owned many medium-sized estates, rounded off by occupation of public land; but there is little evidence of consolidated plantation-type estates such as were found in some provinces and later in Italy as well. Archaeological evidence, even quite near Rome, predominantly shows small- and medium-scale farming. This, however, might well go with the substitution of slaves and tenants (*coloni*) for free owners and the drift to the towns on the part of the latter.

The reform movement of the Gracchi (133–121 BC). From the state's point of view, the chief effect was a decline in military manpower. The minimum property qualification for service was lowered, the minimum age (17) ignored; resistance became frequent, especially to the distant and unending guerrilla war in Spain.

The program and career of Tiberius Sempronius Gracchus. Tiberius Gracchus, grandson of Scipio Africanus and son of the Gracchus who had conquered the Celtiberi and treated them well, was quaestor in Mancinus' army when it faced annihilation; on the strength of his family name, he personally negotiated the peace that saved it. When the Senate—on the motion of his cousin Scipio Aemilianus, who later finished the war—renounced the peace, Tiberius felt aggrieved and joined a group of senior senators hostile to Aemilianus and with ideas on reform. Elected tribune for 133, in Scipio's absence, Tiberius attempted to find a solution for the social and military crisis, with the political credit to go to himself and his backers. Tiberius had no intention of touching private property; his idea was to enforce the legal but widely ignored limit of about 300 acres (120 hectares) on occupation of public land and to use

Plan to settle landless citizens

the land thus retreived for settling landless citizens, who would both regain a secure living and be liable for service. A slave war in Sicily, lasting several years and threatening to spread to Italy, had underlined both the danger of using large numbers of slaves on the land and the need for a major increase in military citizen manpower.

Tiberius' proposal was bound to meet with opposition in the Senate, which consisted of large landowners. On the advice of his eminent backers, he took his bill—which made various concessions to those asked to obey the law and hand back excess public land—straight to the Assembly of the Plebs, where it found wide support. This was not revolutionary; bills directly concerning the People appear to have been frequently passed in this way. But his opponents persuaded another aristocratic tribune, Marcus Octavius, to veto the bill. Tiberius tried the constitutional riposte: an appeal to the Senate for arbitration. But the Senate was unwilling to help, and Octavius was unwilling to negotiate over his veto—an action apparently unprecedented, though not (strictly speaking) unconstitutional. Tiberius had to improvise a way out of the impasse. He met Octavius' action with a similarly unprecedented retort, and had Octavius deposed by the Assembly. He then passed his bill in a less conciliatory form and had himself, his father-in-law, and his brother appointed commissioners with powers to determine boundaries of public land, confiscate excess acreage, and divide it in inalienable allotments among landless citizens. As it happened, envoys from Pergamum had arrived to inform the Senate that Attalus III had died and made the Roman People his heirs (provided the cities of his kingdom were left free). Tiberius, at whose house the envoys were lodging, anticipated Senate debate and had the inheritance accepted by the People and the money used to finance his agrarian schemes.

Tiberius' opponents now charged him with aiming at tyranny, a charge that many may well have believed: redistribution of land was connected with demagogic tyranny in Hellenistic states, and Tiberius' subsequent actions had been high-handed and beyond the flexible borderline of what was regarded as *mos majorum* (constitutional custom). Fearing prosecution after he left office, he now began to canvass for a second tribunate—another unprecedented act, bound to reinforce fears of tyranny. The elections took place in an atmosphere of violence, with nearly all his tribunician colleagues now opposed to him. When the consul Publius Scaevola, on strict legal grounds, refused to act against him, Publius Scipio Nasica, the chief pontiff, led a number of senators and their clients to the Assembly, and Tiberius was killed in a resulting scuffle. Widespread and bloody repression followed in 132. Thus political murder and political martyrdom were introduced into Roman politics.

Introduction of political murder and martyrdom into Roman politics

The land commission, however, was allowed to continue because it could not easily be stopped. Some evidence of its activities survives. By 129, perhaps running out of available land held by citizens, it began to apply the Gracchan law to public land held by Italian individuals or communities. This had probably not been envisaged by Tiberius, just as he did not include noncitizens among the beneficiaries of distributions. The Senate, on the motion of Scipio Aemilianus, upheld the Italians' protests, transferring decisions concerning Italian-held land from the commission to a consul. This seriously hampered the commission's activities. Marcus Fulvius Flaccus, chairman of the commission and consul in 125, tried to solve the problem by offering the Italians the citizenship (or alternatively the right to appeal against Roman executive acts to the Roman People) in return for bringing their holdings of public land under the Gracchan law. This aroused fears of uncontrollable political repercussions. Flaccus was ordered by the Senate to fight a war in southern France (where he gained a triumph) and had to abandon his proposal. There is no sign of widespread Italian interest in it at this time, though the revolt of the Latin colony Fregellae (destroyed 125) may be connected with its failure.

The program and career of Gaius Sempronius Gracchus. In 123 Gaius Gracchus, a younger brother of Tiberius, became tribune. He had served on Tiberius' land commission and had supported Flaccus' plan. Making the most of his martyred brother's name, Gaius embarked on a scheme of general reform in which, for the first time in Rome, Greek theoretical influences may be traced. Among many reforms—including provision for a stable and cheap wheat price and for the foundation of colonies (one on the site of Carthage), to which Italians were admitted—two major ideas stand out: to increase public revenues (both from the empire and from taxes) and pass the benefit on to the people; and to raise the wealthiest nonsenators (particularly the equites, holders of the "public horse"—who received state financial aid for the purchase and upkeep of their horses—and next to senators in social standing) to a

position from which, without actually taking part in the process of government, they could watch over senatorial administration and make it more responsible. The idea was evoked by Tiberius' death. As early as 129 a law compelled senators to surrender the "public horse" (which hitherto they had also held) and possibly in other ways enhanced the class-consciousness and privileges of the equites. Gaius put the *publicani* (public contractors, hitherto chiefly concerned with army and building contracts and with farming minor taxes) in charge of the main tax of Asia—a rich province formed out of Attalus' inheritance, which would henceforth provide Rome with the major part of its public revenues. This was expected both to reduce senatorial corruption and to improve efficiency. Gaius also put eminent nonsenators (probably defined by wealth, but perhaps limited to the equites, or equestrian class) in charge of the *quaestio repetundarum,* whose senatorial members had shown too much leniency to their colleagues, and he imposed severe penalties on senators convicted by that court. Finally, in a second tribunate, he hoped to give citizenship to Latins and Latin rights to other Italians, with the help of Flaccus who, though a distinguished former consul, took the unique step of becoming tribune. But a consul and a tribune of 122 combined to persuade the citizen voters that it was against their interests to share the privileges of citizenship: the bill was defeated, and Gaius failed in his attempt to be reelected once more. In 121, preparing (as private citizens) to use force to oppose the cancellation of some of their laws, Gaius and Flaccus were killed in a riot, and many of their followers were executed.

During the next decade the measures benefitting the people were largely abolished, though the Gracchan land distributions, converted into private property, did temporarily strengthen the Roman citizen peasantry. The provisions giving power to wealthy nonsenators could not be touched, for political reasons, and they survived as the chief effect of Gaius' tribunates. The court seems to have worked better than before, and, during the next generation, several other standing criminal courts were instituted, as were occasional ad hoc tribunals, always with the same class of jurors. In 106 a law adding senators to the juries was passed, but it remained in force for only a short time.

The republic (c. 121–91 BC). *War against Jugurtha.* Since Roman historians were no more interested in politics than (on the whole) in social or economic developments, the struggles of the aristocratic families must be pieced together from chance information. It would be mere paradox to deny the importance in republican Rome, as in better known aristocratic republics, of family feuds, alliances, and policies, and parts of the picture are known: *e.g.,* the central importance of the family of the Metelli, prominent in politics for a generation after the Gracchi and dominant for part of that time. In foreign affairs the client kingdom of Numidia—loyal ever since its institution by Scipio Africanus—assumed quite unwarranted importance when a succession crisis developed there soon after 120, as a bastard, Jugurtha, relying on superior ability and aristocratic Roman connections, sought to oust his two legitimate brothers from their shares of the divided kingdom. Rome's usual diplomatic methods failed to stop Jugurtha who disposed of his brothers, but the massacre of Italian settlers at Cirta by his soldiers forced the Senate to declare war (112). The war was waged reluctantly and ineffectively, with the result that charges of bribery were freely bandied about by demagogic tribunes taking advantage of suspicion of aristocratic political behaviour that had smouldered ever since the Gracchan crisis. Significantly, some eminent men, hated from those days, were now convicted of corruption. The Metelli, however, emerged unscathed, and Quintus Metellus, consul in 109, was entrusted with the war in Africa. He waged it with obvious competence but failed to finish it and thus gave Gaius Marius, a senior officer, his chance.

Struggles of the aristocratic families

The career of Gaius Marius. Marius, born of an equestrian family at Arpinum, had attracted the attention of Scipio Aemilianus as a young soldier and, by shrewd political opportunism, had risen to the praetorship and married into the patrician family of the Julii Caesares. Though he had deeply offended the Metelli, once his patrons, Marius' military talents were considerable, and Quintus Metellus had taken him to Africa as a *legatus.* Marius intrigued against his commander in order to gain a consulship; he was elected (chiefly with the help of the equites and anti-aristocratic tribunes) for 107 and was given charge of the war by special vote of the People. He did little better than Metellus had, but in 105 his quaestor Lucius Sulla, in delicate and dangerous negotiations, brought about the capture of Jugurtha, opportunely winning the war for Marius and Rome.

During the preceding decade a serious threat to Italy had developed in the north. Starting in 125, several Roman commanders (Marcus Flaccus has been noted) had fought against Ligurian and Gallic tribes in southern France and had finally established a Roman sphere of influence there: a road had been built linking Italy with Spain, and some garrison posts probably secured it; finally, a colony was settled at Narbonne, an important road junction (*c.* 118). But unwilling to extend administrative responsibilities, the Senate had refused to establish a regular *provincia.* Then some migrating German tribes, chief of them the Cimbri, after defeating a Roman consul, invaded southern France, attracting native sympathy and finding little effective Roman opposition. Two more consular armies suffered defeat, and in October 105 a consul and proconsul with their forces were destroyed at Orange. There was panic in Rome, allayed only by the firm action of the other consul, Publius Rutilius Rufus.

Roman sphere of influence in southern France

At this moment news of Marius' success in Africa arrived, and he was at once dispensed from legal restrictions and again elected consul for 104. After a brilliant triumph, restoring Roman morale, he took over the army prepared and trained by Rutilius and was reelected consul year after year, while the German tribes delayed attacking Italy. Finally, in 102–101, he annihilated them at Aquae Sextiae (Aix-les-Bains) and, with his colleague Quintus Catulus, on the Campi Raudii (near the Po Delta). Another triumph and a sixth consulship (in 101) were his reward.

In his first consulship, Marius had taken a step of great (and probably unrecognized) importance: aware of the difficulties long endemic in the traditional system of recruitment, he had ignored property qualifications in enrolling his army and, as a result, had recruited ample volunteers from among men who had nothing to lose. This radical solution was thenceforth generally imitated, and conscription became confined to emergencies (like the Social and Civil wars). He also enhanced the importance of the legionary eagle (standard), thus beginning the process that led to each legion's having a continuing corporate identity. At the same time, Rutilius introduced arms drill and reformed the selection of senior officers. Various tactical reforms in due course led to the increasing prominence of the cohort (one tenth of a legion) as a tactical unit and the total reliance on non-Roman auxiliaries for light-armed and cavalry service. The precise development of these reforms cannot be traced, but they culminated in the much more effective armies of Pompey and Caesar.

Marius' African army had been unwilling to engage in another war, and Marius preferred to use newly levied soldiers (no longer difficult to find). But neither he nor the Senate seemed aware of any responsibilities to the veterans. In 103 a tribune, Lucius Saturninus, offered to pass a law providing land in Africa for them in return for Marius' support for some anti-oligarchic activities of his own. Marius agreed, and the large lots distributed to his veterans (both Roman and Italian) turned out to be the beginning of the Romanization of Africa. In 100, with the German wars ended, Saturninus again proved a welcome ally, arranging for the settlement of Marius' veterans in Gaul. An incidental effect was the departure of Marius' old commander and subsequent enemy, Quintus Metellus, who refused to recognize the validity of Saturninus' law and, choosing martyrdom, went into exile. But this time Saturninus exacted a high price. With his ally, the praetor Gaius Glaucia, he introduced laws to gain the favour of plebs and equites and proceeded to provide for the settlement of veterans of wars in Macedonia and Sicily in the same way as for those of Marius' war. He planned to

seek reelection for 99, with Glaucia illegally gaining the consulship. Violence and even murder were freely used to accomplish these aims.

Marius' opposition to Saturninus and Glaucia

Marius now had to make a choice. Saturninus and Glaucia might secure him the continuing favour of the plebs and perhaps the equites, though they might also steal it for themselves. But as the saviour of his country, six times consul, he now hoped to become an elder statesman (*princeps*), accepted and honoured by those who had once looked down on him as an upstart. To this end he had long laboured, dealing out favours to aristocrats who might make useful allies. This was the only accepted reward for achievement; Marius never thought of revolution or tyranny. Hence, when called on to save the state from his revolutionary allies, he could not refuse. He imprisoned them and their armed adherents and did not prevent their being lynched. Yet, having saved the oligarchy from revolution, he received little reward; he lost the favour of the plebs, while the oligarchs, in view of both his birth and his earlier unscrupulous ambition, refused to accept him as their equal. Metellus was recalled; this was a bitter blow to Marius' prestige, and he preferred to leave Rome and visit Asia.

Before long a face-saving compromise was found, and Marius returned; but in the 90s he played no major part. Though he held his own when his friends and clients were attacked in the courts, his old aristocratic protégés now found more promising allies. Sulla is typical: closely associated with Marius in his early career, he was by 91 ready to take the lead in attacking Marius and (significantly) found eager support. The oligarchy could not forgive Marius.

Wars and dictatorship (c. 91–80 BC). *Events in Asia.* In foreign affairs the 90s were dominated by Asia, Rome's chief source of income. Mithradates VI, king of Pontus, had built a large empire around the Black Sea and was probing and intriguing in the Roman sphere of influence. Marius had met him and had given him a firm warning, temporarily effective: Mithradates had proper respect for Roman power. Scheming to annex Cappadocia, he had been thwarted by the Senate's instructing Sulla, as proconsul, to install a pro-Roman king there in 96–95. (It was on this occasion that Sulla received a Parthian embassy—the first contact between the two powers.) But dissatisfaction in the Roman province gave new hope to Mithradates. Ineffectively organized after annexation and corrupt in its cities' internal administration, Asia was soon overrun with Italian businessmen and Roman tax collectors. When the Senate realized the danger, it sent its most distinguished jurist, Quintus Mucius Scaevola (consul in 95 and *pontifex maximus*), on an unprecedented mission to reorganize Asia (94). He took Publius Rutilius Rufus—jurist, Stoic philosopher, and former consul—with him as his senior officer, and after Scaevola's return Rutilius remained behind, firmly applying the new principles. This caused an outcry from businessmen, whose profits he had kept within bounds: he was prosecuted for "extortion" in 92 and convicted after a trial in which Roman *publicani* and businessmen unscrupulously used their power among the class that provided criminal juries. The verdict revealed the breakdown of Gaius Gracchus' system: the class he had raised to watch over the Senate now held irresponsible power, making orderly administration impossible and endangering the empire. Various leading senators were at once vexatiously prosecuted, and political chaos threatened.

Developments in Italy. The 90s also saw dangerous developments in Italy. In the 2nd century BC, Italians as a whole had shown little desire for Roman citizenship and had been remarkably submissive under exploitation and ill-treatment. The most active of their governing class flourished in overseas business, and the more traditionally minded were content to have their oligarchic rule supported by Rome. Their admission to citizenship had been proposed as a by-product of the Gracchan reforms. By 122, it had become clear that the Roman people agreed with the oligarchy in rejecting it. The sacrifices demanded of Italy in the Numidian and German wars probably increased dissatisfaction among Italians with their patent inferiority. Marius gave citizenship to some as a reward for military distinction—illegally, but his standing (*auctoritas*) sufficed to defend his actions. Saturninus admitted Italians to veteran settlements and tried to gain citizenship for some by full admission to Roman colonies. The censors of 97–96, aristocrats connected with Marius, shared his ideas and freely placed eminent Italians on the citizen registers. This might have allayed dissatisfaction, but the consuls of 95 passed a law purging the rolls and providing penalties for those guilty of fraudulent arrogation. The result was insecurity and danger for many leading Italians. By 92 there was talk of violence and conspiracy among desperate men.

Insecurity and danger for leading Italians

It was in these circumstances that the eminent young noble, Marcus Livius Drusus, became tribune for 91 and hoped to solve the menacing accumulation of problems by means of a major scheme of reforms. He attracted the support of the poor by agrarian and colonial legislation and tried to have all Italians admitted to citizenship and to solve the jury problem by a compromise: the courts would be transferred to the Senate, and 300 equites would be admitted to it. (To cope with the increase in business it would need this increase in size.) Some leading senators, frightened at the dangerous situation that had developed, gave weighty support. Had Drusus succeeded, the poor and the Italians might have been satisfied; the equities, deprived of their most ambitious element by promotion, might have acquiesced; and the Senate, always governed by the prestige of the noble *principes* rather than by votes and divisions, could have returned, little changed by the infusion of new blood, to its leading position in the process of government. But Drusus failed. Some members of each class affected were more conscious of the loss than of the gain; and an active consul, Lucius Philippus, provided leadership for their disparate opposition. After much violence, Drusus' laws were declared invalid. Finally he himself was assassinated. The Italians now rose in revolt (the Social War), and in Rome a special tribunal, manned by the Gracchan jury class, convicted many of Drusus' supporters until the Senate succeeded in suspending its sittings because of the military danger.

The Social War

The first year of the Social War (90) was dangerous: the tribes of central and southern Italy, traditionally among the best soldiers in Rome's wars, organized in a confederacy for the struggle that had been forced upon them. Fortunately all but one of the Latin cities—related to Rome by blood and tradition and specially favoured by Roman law—remained loyal: their governing class had for some time had the privilege of automatically acquiring Roman citizenship by holding local office. Moreover, Rome now showed its old ability to act quickly and wisely in emergencies: the consul Lucius Caesar passed a law giving citizenship to all Italians who wanted it. The measure came in time to head off major revolts in Umbria and Etruria, which accepted at once.

Civil war and the rule of Lucius Sulla. In 89 the war in central Italy was won, and Gnaeus Pompeius Strabo celebrated a triumph. Attention now turned to the East, where Mithradates had taken advantage of Rome's troubles to expel the kings of Cappadocia and Bithynia. A Roman embassy restored them, and he withdrew; but, when the envoys incited Bithynian incursions into his territory, Mithradates launched a major offensive, overrunning the two kingdoms and invading Roman territory, where he attracted the sympathy of the natives by executing thousands of Italians and defeated and captured the Roman commanders in the area.

In Rome, various men, including Marius, had hoped for the Eastern command. But it went to Sulla, elected consul for 88 after distinguished service in the Social War. Publius Sulpicius, a tribune in that year and an old friend of Drusus, tried to continue the latter's policy of justice to the Italians by abolishing the gerrymandering that in practice deprived the new citizens of an effective vote. Finding the oligarchy firmly opposed, he gained the support of Marius (who still commanded much loyalty) for his plans by having the Eastern command transferred to him. After much street-fighting, the consuls escaped from Rome, and Sulpicius' bills were passed. Sulla's response was totally unforeseen: appealing to the army he had led

Sulla's occupation of Rome

in the Social War, still engaged in mopping-up operations in Campania, he persuaded them to march on Rome, which he occupied, executing Sulpicius; Marius and others escaped. Significantly, Sulla's officers left him. It was the first time a private army of citizens had occupied Rome—an effect of Marius' army reform, which had ended by creating a "client army" loyal chiefly to its commander, and of the Social War, which had made the use of force within Italy seem commonplace. The end of the republic was foreshadowed.

Having cowed Rome into acquiescence and passed some legislation, Sulla left for the East. Cinna, one of the consuls of 87, at once called for the overthrow of Sulla's measures. Resisted by his colleague Octavius, he left Rome to collect an army and, with the help of Marius, occupied the city after a siege. Several leading men were killed or condemned to death, Sulla and his supporters were outlawed, and (after Marius' death early in 86) another commander was sent to Asia. The policy now changed to one of reconciliation: the Social War was wound up, and the government gained wide acceptance until Cinna was killed by mutinous soldiers (84).

Sulla meanwhile easily defeated Mithradates' forces in two battles in Boeotia, took Athens, which under a revolutionary regime had declared for Mithradates, and cleared the King's army out of Greece. While negotiating with Cinna's government, Sulla also entered upon negotiations with Mithradates and, when he heard of Cinna's death, quickly made peace and an alliance with Mithradates, driving the government's commander in Asia to suicide. After wintering his troops in the rich cities of Asia, Sulla crossed to Greece and then to Italy, where his veteran army broke all resistance and occupied Rome (82). Sulla was elected dictator and, while Italy and all the provinces except Spain were quickly reduced, began a reign of terror (the "proscriptions"), in which hundreds of his enemies or those of his adherents were killed without trial, while their property went to enrich him and his friends. Wherever he had met resistance in Italy, land was expropriated and given to his soldiers for settlement.

While the terror was going on, Sulla used his powers to put through a comprehensive program of reform (81). Although he had twice taken Rome with a private proletarian army, he had earlier had connections with the inner circles of the oligarchy, and after Cinna's death some eminent men who had refused to collaborate with Cinna joined Sulla. By the time Sulla's success seemed certain, most of those who had collaborated were also on his side, and he was acclaimed as the defender of the nobility who had defeated an illegal revolutionary regime. His reforms aimed chiefly at stabilizing Senate authority by removing alternative centres of power. The tribunate was emasculated; the censors' powers were reduced; provincial governors were subjected to stricter Senate control; and the equites, who had been purged of Sulla's opponents by the proscriptions, were deprived of some symbols of dignity and made leaderless by the inclusion of 300 of Sulla's chief supporters in the Senate. The jury reform of Gaius Gracchus, seen by some leading senators as the prime cause of political disintegration, could now be undone and the criminal courts could once more become a monopoly of senators.

Sulla's attempt to restore stable government

Sulla's measures were by no means merely reactionary. His program was basically that of Marcus Drusus. His overriding aim was the restoration of stable government, and this could only come from the Senate, directed by the *principes* (former consuls and those they chose to consult). He accepted and even extended recent developments where they seemed useful: the Italians retained full citizenship; the system of standing criminal courts was expanded to include six or seven major crimes; the practice of praetors normally spending their year of office in Rome and then going to provinces for a second year was extended to consuls and became an integral part of his system. To prevent long command of armies (which might lead to careers like his own), he increased the number of praetors so that, in principle and in normal circumstances, each province might have a new governor every year. As for the overriding problem of poverty, his contribution to solving it was to settle tens of thousands of his veterans on land confiscated from enemies in Italy; they would be turned into "haves" and would be ready (using their military training) to defend the social order, in which they now had a stake, against "have-nots"—both old and newly created by proscription and confiscation.

At the beginning of 80, Sulla laid down his dictatorship and became merely consul, with the senior Metellus (Quintus Metellus Pius), a relative of his wife, as his colleague. The state of emergency was officially ended. At the end of the year, Sulla retired to Campania as a private citizen, after seeing to the election of two reliable consuls, hoping that the restored oligarchy would learn to govern the state he had handed over to them. For 78 Marcus Lepidus, an ambitious patrician whom Sulla disliked and distrusted, was elected consul. Sulla did not intervene. Within a few months, Sulla was dead. Lepidus at once attacked his system, using the grievances of the expropriated as a rallying cry and his province of Gaul as a base. But he was easily defeated by his former colleague Quintus Catulus, assisted by young Gnaeus Pompeius (Pompey).

The Roman state in the two decades after Sulla (79–60 BC). *The early career of Pompey.* Pompey was the son of Gnaeus Pompeius Strabo, who had triumphed after the Social War but had incurred general hatred because of cold-blooded duplicity during the troubles of 88 and 87. After Strabo's death, young Pompey, who had served under him and inherited his dubiously won wealth, was protected by Cinna's government against his father's enemies. Following in his father's footsteps, he deserted the government after Cinna's death, raised a force among his father's veterans in central Italy, and helped to conquer Italy and, in a lightning campaign, Sicily and the province of Africa for Sulla. Though not old enough to hold any regular magistracy (he was born in 106), he had, from these military bases, blackmailed Sulla into granting him a triumph (81) and had married into the core of the Sullan oligarchy. Out of pique against Sulla, he had supported Lepidus' election for 78, but he had too great a stake in the Sullan system to permit Lepidus to overthrow it.

Revolt of Sertorius

Meanwhile a more serious challenge to the system had arisen in Iberia. Quintus Sertorius, a former praetor of tough Sabine gentry stock, had refused to follow most of his social betters in joining the invader of Italy; he had left for Spain, where he claimed to represent the legitimate government. Although acting throughout as a Roman proconsul, with a "counter-Senate" of eminent Roman citizens, Sertorius won the enthusiastic support of the natives by his fairness, honesty, and charisma, and he soon held most of the Iberian Peninsula, defending it successfully even against a large force under Quintus Metellus Pius. The consuls of 77 would have nothing to do with this war, and so Pompey was entrusted by the Senate, through the efforts of his eminent friends and sponsors, with the task of assisting Metellus. The war dragged on for years, with little glory for the Roman commanders. Sertorius had many sympathizers in Italy; but superior numbers and resources finally wore him down, and he was assassinated by a Roman officer. Pompey easily defeated the remnants of his forces in 72.

The death of Nicomedes IV of Bithynia (74) led to another major war. Like Attalus of Pergamum, Nicomedes left his kingdom to Rome, and this provoked Mithradates, who was in contact with Sertorius and knew of Rome's difficulties, to challenge Rome again. The Eastern command again led to intrigues in Rome, where pressure against the Sullan system had brought a relaxation of the restrictions on the tribunate; amid increasing agitation, the armies in Spain (especially Pompey's) had come to appear as future arbiters of Roman politics. The command finally went to Lucius Lucullus, a relative of Sulla and consul in 74, who hoped to build up a countervailing power in the East.

At the same time, Marcus Antonius, father of the later Triumvir, was given a command against the pirates in the eastern Mediterranean (whom his father had already fought in 102–100), partly, perhaps, as further re-insurance against Pompey. With Italian manpower heavily committed, a minor slave rising led by Spartacus (73) assumed threatening dimensions, until Marcus Crassus (an old

Sullan and profiteer in the proscriptions) volunteered to accept a special command and defeated the slaves. At this point (71), Pompey arrived back from Spain with his army, crucified the remnants of the slave army, and claimed credit for the victory.

Pompey and Crassus. He and Crassus now confronted each other, each demanding the consulship for 70, though Pompey had held no regular magistracy and was not a senator. Agreeing to join forces, both secured it.

During their consulship, the political, though not the administrative, part of the Sullan settlement was repealed. The tribunes' powers were fully restored, criminal juries were divided between senators and wealthy nonsenators, and, for the first time since Sulla, censors—both supporters of Pompey—were elected, purged the Senate, and in compiling the registers at last fully implemented the Italians' citizenship. The year 70 also saw the prosecution of Verres (son of a "new man" and Sullan profiteer), who had surpassed the very liberal Roman conventions in exploiting his province of Sicily, relying for future impunity on his aristocratic connections (especially the Metelli and their friends), his fortune, and the known corruptibility of the Sullan senatorial juries. But Verres was unlucky. First, he had ill-treated some of Pompey's important Sicilian clients, thus incurring Pompey's displeasure; next, his case coincided with the anti-Sullan reaction of 70; finally, the Sicilians succeeded in persuading Cicero—an ambitious young "new man" from Arpinum, hoping to imitate the success of his fellow citizen Marius by means of his rhetorical ability—to undertake the prosecution. Despite obstruction from Verres' friends, Cicero collected massive evidence against him, presented his case to fit into the political context of the year, and obtained Verres' conviction as an act of expiation for the shortcomings of the Sullan order.

Loss of control by the Sullan establishment

The year 70 thus marks the loss of control by the Sullan establishment. The nobility (families descended from consuls) continued to gain most of the consulships, with the old patriciate (revived by Sulla after a long decline) stronger than for generations; the Senate still supervised administration and took ordinary political decisions; the system continued to rely essentially on *mos majorum* (constitutional custom) and *auctoritas* (prestige)—potent forces in the status society of the Roman Republic. But the solid bases of law and power that Sulla had tried to give it had been surrendered. The demagogue—tribune or consul—could use the legal machinery of the popular assembly (hence such men are called *populares*), while the commander could rely on his army, in the pursuit of private ambition. The situation that Sulla had tried to remedy now returned, made worse by his intervention. His massacres and proscriptions had weeded out the defenders of lawful government, and his rewards had gone to the timeservers and the unscrupulous. The large infusion of equites into the Senate had intensified the effect. While eliminating the serious friction between the two classes, which had made the state ungovernable by 91, it had filled the Senate with men whose tradition was the opposite of that sense of mission and public service that had animated the best of the aristocracy. Few men in the new ruling class saw beyond self-interest and self-indulgence.

One result was that massive bribery and civil disorder in the service of ambition became endemic. Laws were repeatedly passed to stop them, but they remained ineffective because few found it in their interest to enforce them. Exploitation of the provinces did not decrease after Verres: governors (still with unlimited powers) feathered their own nests and were expected to provide for all their friends. Extortion cases became a political ritual, with convictions impossible to obtain. Cicero, thenceforth usually counsel for the defence, presents hair-raising behaviour as commonplace and claims it as acceptable. The Senate's traditional opposition to annexation faded out. Pompey made Syria into a province and added a large part of Pontus to Bithynia (inherited in 74 and occupied in 70); the demagogue Clodius annexed Cyprus—driving its king to suicide—to pay for his massive grain distributions in Rome; Caesar, finally, conquered Gaul by open aggression and genocide and bled it white for the benefit of his friends and his ambitions. Crassus would have done the same with Parthia, had he succeeded. Opposition to all this in the Senate, where it appeared, was based on personal or political antagonism. If the robber barons were attacked on moral grounds, it was because of the use they made of their power in Rome.

Politically, the 60s lay under the shadow of Pompey. Refusing to take an ordinary province in 69, he waited for his chance. It came in 67, when his adherent Gabinius, as tribune, secured him, against the opposition of all important men, an extraordinary command, with unprecedented powers, to deal with the pirates. Pompey succeeded within a few months where Antonius and others had failed. The equites and the people were delighted because trade, including Rome's food imports, would now be secure. Meanwhile Lucullus had driven Mithradates out of Asia Minor and into Armenia; but he had offended Roman businessmen by strict control and his own soldiers and officers by strict discipline. Faced with mutinies, he suffered a reverse and became vulnerable to attacks in Rome. In 66 another tribunician law appointed Pompey, fresh from his naval victories, to take over supreme command in the East, which he did at once, studiously insulting his predecessor. He quickly defeated Mithradates and procured his death, then spent some time in a total reorganization of the East, where Asia (the chief source of revenue) was protected by three further provinces and a ring of client states beyond the frontier. The whole of the East now stood in his clientela, and most of it owed him money as well. He returned as by far the wealthiest man in Rome.

Defeat and death of Mithradates VI

Political suspicion and violence. Meanwhile Roman politics had been full of suspicion and violence, much of it stirred up by Crassus who, remembering 71, feared Pompey's return and tried to make his own power impregnable. There was much material for revolution, with poverty (especially in the country, among families dispossessed by Sulla) and debt (among both the poor and the dissolute rich) providing suitable issues for unscrupulous *populares*.

One such man, the patrician Catiline, after twice failing to gain the consulship by traditional bribery and intrigue, put himself at the head of a movement planning a coup d'état in Rome to coincide with an armed rising in Italy (late 63). Cicero, as consul, defeated these efforts and, relying on the doubtful legality of a Senate vote in support, had Catiline's eminent Roman associates executed. Catiline himself fell in a desperate battle.

Isolation of Cicero

For Cicero—the "new man" who had made his way to the top by his own oratorical and political skill, obliging everyone by unstinting service, representing Pompey's interests in Rome while avoiding offense to Pompey's enemies—this was the climax of his life. Like his compatriot Marius, he had saved the state for its rulers: he had taken resolute action when those rulers were weak and vacillating; and, like Marius, he got small thanks for it. Pompey was miffed at having to share his fame with a municipal upstart, and eminent gentlemen could not forgive that upstart for having driven patricians to their death.

Pompey's return was peaceful. Like Marius, he wanted recognition, not tyranny. He dismissed his army, to the surprise of Crassus and others, and basked in the glory of his triumph and the honours voted to him. But having given up power, he found himself caught in a net of constitutional obstruction woven by his politically experienced enemies and was unable to obtain either of his principal demands: land for his veterans and the ratification of his arrangements in the East. It was at this point that Caesar returned from Spain.

Gaius Julius Caesar, descended (as he insisted) from kings and gods, had shown talent and ambition in his youth: he opposed Sulla, but without inviting punishment; married into the oligarchy, but advocated popular causes; vocally defended Pompey's interests, while aiding Crassus in his intrigues and borrowing a fortune from him; flirted with Catiline, but refused to dabble in revolution, then worked to save those whom Cicero executed. In 63 he won a startling success: defeating two distinguished *principes,* he, who had not yet been praetor, was elected *pontifex maximus*—a post of supreme dignity, power, and patronage. The Roman aristocracy had long ceased to believe in the

state religion; but its ceremonial was kept up for popular effect and as a recognized means of political manipulation, so that priesthoods could give more lasting power than magistracies, in addition to the cachet of social success. Young Caesar was now head of the hierarchy. After his praetorship (62), Caesar successfully governed Spain, clearing a surplus sufficient to pay off his debts. On returning to Rome, he naturally hoped for the consulship of 59; but his enemies, by legal chicanery, forced him to choose between standing for office and celebrating a triumph. He gave up the triumph and easily became consul.

Alliance of Caesar, Pompey, and Crassus

The final collapse of the Roman republic (59–44 BC). *Caesar, Pompey, and Crassus.* For his consulship Caesar fashioned an improbable alliance: his skill in having won the trust of both Crassus and Pompey enabled him to unite these two enemies in his support. Crassus had the connections, Pompey had the soldiers' vote, and Caesar was consul and *pontifex maximus.* The combination (often misleadingly called the "first Triumvirate") was invincible, especially since the consul Caesar had no scruples about countering legal obstruction with open force. Pompey got what he wanted, and so did Crassus (whose immediate need was a concession to the Asian tax farmers, in whose companies he probably had much of his capital). In return, Caesar got a special command in Cisalpine Gaul and Illyricum for five years by vote of the People; the Senate itself, on Pompey's motion, extended it to Transalpine Gaul. Marriage alliances sealed the compact, chief of them Pompey's marriage to Caesar's daughter Julia.

Caesar left for Gaul, but Rome was never the same; the shadow of the alliance hung over it, making the old-style politics impossible. In 58 Publius Clodius, another aristocratic demagogue, was tribune and defended Caesar's interests. Cicero had incurred Clodius' enmity and was now sacrificed to him: he was driven into exile as having unlawfully executed citizens in 63. By 57 Caesar's allies had drifted back into rivalry. Pompey secured Cicero's return, and Cicero at once tried to break up the alliance by attracting Pompey to the Senate's side. Just when he seemed about to succeed, the three dynasts secretly met and revived their compact (56). Rome had to bow once more. In 55 Pompey and Crassus were consuls, and the contents of their secret agreement were slowly revealed. Caesar, whom his enemies had made efforts to recall, was prolonged in his command for five years and (it later appeared) had been promised another consulship straight after, to secure him against prosecution and give him a chance of another army command. Pompey was given a special command over all of Spain, which he exercised through deputies while he himself remained just outside Rome to keep an eye on the city. Crassus, who now needed glory and new wealth to equal those of his allies, was to attack Parthia with a large army. Thus the three dynasts would practically monopolize military power for the foreseeable future.

Cicero, among others, had to submit and was thenceforth their loyal spokesman. After his achievement of 63 he had dreamt of leading a coalition of all "right-thinking" men in Italy in defending the traditional oligarchy, but he had found little support among the oligarchy and now used this fact to rationalize his surrender. His brother took service in Gaul under Caesar.

The dynasts' pact did not even bring peace. Clodius, as tribune, had created a private army, and there was no state force to counter it. Pompey could have done it by calling his soldiers in, but the Senate did not trust him enough to request this, and Pompey did not wish to parade himself as an unashamed tyrant. Other men formed private armies in opposition to Clodius, and one Milo at last managed to have him killed after a scuffle (52). By then, however, Roman politics had radically and unexpectedly changed.

End of the pact

Political manoeuvres. Julia died in 54, breaking the ties between Caesar and Pompey. Caesar pressed Pompey to renew them, but Pompey held off, preserving his freedom of action. Crassus' Parthian campaign ended in disaster and in Crassus' death (53). By 52 Pompey and Caesar stood face to face, still nominally friends but with no personal link between them and no common interests. Caesar, by conquering the whole of Gaul, had almost equalled Pompey's prestige and, by his utterly ruthless way of waging war, Pompey's wealth. Unlike Pompey, he used his wealth to dispense patronage and buy useful friends. At this point Pompey cautiously offered the oligarchy his support. It had much to give him that he wanted—control of the administrative machine, respectability, and the seal of public approval. Its leaders (even the intransigent young Cato, who had led opposition to the three individually long before their alliance and to their joint oppression of the state ever since) now recognized that acceptance of Pompey's terms and surrender to his protection was their only chance of survival. Pompey at once turned firmly against Milo, who presented a political threat: if Milo could use the force that had killed Clodius to keep firm control of Rome, he—an ambitious man of known conservative views—might in due course offer an alternative and more trustworthy champion to the oligarchy. But he was not yet ready. Pompey forced them to make their choice at once, and they chose Pompey in preference. He was made sole consul and had Milo convicted by an intimidated court. Meanwhile he had made a marriage alliance with the noblest man in Rome, Quintus Metellus Scipio, who became his colleague in the consulship. The state had captured Pompey (or vice versa), and Caesar stood alone in opposition to both of them. During the next two years there were a series of manoeuvres: the Senate leaders, with Pompey's silent support, worked for Caesar's recall, which would have meant his instantly sharing the fate of Milo; while Caesar and his agents in Rome tried to strike some bargain that would ensure his safety and his future in politics. Finally, Pompey declared himself, and, early in 49, the Senate voted to outlaw Caesar. Two tribunes supporting him (one of them Mark Antony) had to flee. By the time they reached him, Caesar had already crossed the Rubicon: he now had a cause.

Civil war. Pompey had exuded confidence over the outcome if it came to war. In fact, however, Caesar's veterans were unbeatable, and both men knew it. To the disgust of his followers, Pompey evacuated Rome, then Italy. His plan was to bottle Caesar up there and starve him out. But Caesar, in a lightning sweep, seized Massilia and Spain from Pompey's commanders, then crossed to Greece, where a short campaign ended in Pompey's decisive defeat at Pharsalus (48). Pompey fled to Egypt, where he was assassinated by a man hoping thus to curry Caesar's favour. This was by no means the end of the war. Almost at once, Caesar was nearly trapped at Alexandria, where he had intervened in a succession dispute; but he escaped and installed Cleopatra on the throne, for personal as well as political reasons. In Africa the Pompeian forces and their native allies were not defeated until Caesar himself moved against them and annihilated them at Thapsus. Cato, disdaining the victor's pardon, committed suicide at Utica (46). In Spain, where Pompey's name was still powerful, his sons organized a major rising, which Caesar himself again had to defeat at Munda (45) in the bloodiest battle of the war. By the time he returned, he had only a few months to live.

The dictatorship and assassination of Caesar. In Rome the administrative machine had inevitably been disrupted, and Caesar had always remained in control, as consul or as dictator. Those who had feared proscriptions, or hoped for them, were proved wrong. Some of Caesar's enemies had their property confiscated, but it was sold at fair value; most were pardoned and suffered no loss. One of these was Cicero, who, after much soul-searching, had followed his conscience by joining Pompey before Pharsalus.

Changes under Caesar

Poverty and indebtedness were alleviated, but there was no wholesale cancellation of debts or redistribution of property, and many of Caesar's adherents were disappointed. Nor was there a general reform of the republic. (Caesar's only major reform was of the calendar: indeed, the Julian calendar proved adequate for centuries.) The number of senators and magistrates was increased, the citizenship was more freely given, the province of Asia was relieved of some of its tax burden. But Caesar had no plan for reforming the system—not even to the extent that Sulla had tried to do, for Sulla had at least planned for his own

retirement. For a time, honourable men, such as Cicero, hoped that the "Dictator for settling the Constitution" (as Caesar called himself) would produce a real constitution—some return to free institutions. By late 45 that hope was dead. Caesar was everywhere, doing everything to an almost superhuman degree. He had no solution for the crisis of the republic except to embody it in himself and none at all for the hatred of his peers, which he knew this was causing. He began to accept more and more of the honours that a subservient Senate invidiously offered, until finally he reached a position perilously close to kingship (an accursed term in Rome) and even deification. Whether he passed those hazy boundary lines is much debated and not very important. He had put himself in a position in which no Roman ought to have been and which no Roman aristocrat could tolerate. As a loyal friend of his was later to say: "With all his genius, he saw no way out." To escape the problem or postpone it, he prepared for a Parthian war to avenge Crassus—a project most likely to have ended in similar disaster. Before he could start on it, about 60 men—former friends and old enemies, honourable patriots and men with grievances—struck him down in the Senate on March 15, 44 BC.

The Triumvirate and Octavian's achievement of sole power. Brutus and Cassius, the organizers of the conspiracy, expected all Romans to rejoice with them in the rebirth of "freedom." But to the Roman people the freedom of the governing class had never meant very much; the armies (especially in the west) were attached to Caesar; and the Senate was full of Caesarians at all levels, cowed but biding their time. Mark Antony, the surviving consul, whom Brutus had been too scrupulous to assassinate with his master, gradually gained control of the city and the official machinery, and the "liberators" withdrew to the East. But a challenger for the position of leader of the Caesarians soon appeared in the person of Octavian, Caesar's son by adoption and now his heir. Though not yet 20, Octavian proved an accomplished politician, attracting loyalty as a Caesarian while cooperating against Antony with the Senate, which, under Cicero's vigorous leadership, now turned against the consul. Cicero hoped to fragment and thus defeat the Caesarian party, with the help of Brutus and Cassius, who were making good progress in seizing control of the eastern provinces and armies. In 43 the two consuls (both old Caesarian officers) and Octavian defeated Antony at Mutina, and success seemed imminent. But the consuls died, and Octavian demanded and, by armed force, obtained the consulship; and the armies of Italy, Spain, and Gaul soon showed that they would not fight against one another. Octavian, Antony, and Lepidus (the senior Caesarian with an army) now had themselves appointed "Triumvirs for settling the Constitution" for five years and secured control of Italy by massive proscriptions and confiscations (Cicero, Antony's chief enemy, was among the first to die). They then defeated and killed Brutus and Cassius at Philippi (42) and divided the Roman world among themselves, with Lepidus, a weak man accidentally thrust into prominence, getting the smallest share. Octavian, who was to control Italy, met armed opposition from Antony's brother and wife, but they got no help from Antony and were defeated at Perusia (41). Octavian and Antony sealed their alliance with a marriage compact: Antony married Octavia, Octavian's sister. Octavian then confronted Pompey's son Sextus Pompeius, who had seized control of the islands off Italy. After much diplomatic manoeuvring (including another meeting with Antony), Octavian attacked and defeated Sextus; when Lepidus tried to reassert himself, Octavian crushed him and stripped him of his office of Triumvir (while with conspicuous piety leaving him the chief pontificate, now powerless). Octavian now controlled the West and Antony the East, still officially as Triumvirs (their term of office had been extended), even though Lepidus had been eliminated in 36.

Formation of the Triumvirate

Each of the two leaders embarked on campaigns and reorganization in his half—Octavian in Illyricum, Antony particularly on the Parthian frontier. But Antony now married Cleopatra and tried to make Egypt his military and political base. In a war of propaganda, Octavian gradually convinced the western provinces, Italy, and most of the Roman upper class that Antony was sacrificing Roman interests, and trying to become a Hellenistic king in Alexandria, and planning to rule the Roman world from there with Cleopatra. In 32, though he now held no legal position, Octavian intimidated most of Antony's remaining aristocratic friends into joining him, made the whole West swear allegiance to himself, and in 31, as consul, crossed to Greece to attack Antony. On September 2 he defeated Antony and Cleopatra in a naval battle at Actium. Though in itself not a major victory, it was followed by the disintegration of Antony's forces, and Antony and Cleopatra finally committed suicide in Alexandria (30).

(E.Ba.)

THE EARLY ROMAN EMPIRE (31 BC–AD 193)

The consolidation of the empire under the Julio-Claudians. *The establishment of the principate under Augustus.* Actium left Octavian the master of the Roman world. This supremacy, unremittingly maintained until his death more than 40 years later, made him the first of the Roman emperors. Suicide removed Antony and Cleopatra and their potential menace in 30 BC, and the annexation of Egypt with its Ptolemaic treasure brought financial independence. With these reassurances Octavian could begin the task of reconstruction.

The first of the Roman emperors

Law and order had vanished from the Roman state when its ruling aristocrats refused to curb their individual ambitions and thus transformed a republic based on disciplined liberty into a turbulent cockpit of murderous rivalries. Good government depended on limits being set to unrestrained aspirations, and Octavian was in a position to impose them. But, although his military might in 31 BC could guarantee orderly political processes, it was itself incompatible with them; nor did he relish the role of military despot. The fate of Julius Caesar, an eagerness to acquire political respectability, and his own esteem for ancestral custom combined to dissuade Octavian from it. He wished to be, in his own words, "the author of the best civilian government possible." His problem was to regularize his own position so as to make it generally acceptable, without simultaneously reopening the door to violent lawlessness. His pragmatic solution not only ensured stability and continuity, but it also respected republican forms and traditions so far as possible.

Large-scale demobilization allayed men's fears; regular consular elections raised their hopes. In 29–28 BC he carried out, with Agrippa, the first census of the Roman people since 70; and this involved drawing up an electoral roll for the Centuriate Assembly. Elections followed, and Octavian was inevitably chosen consul. Then, on January 13, 27 BC, he offered to lay down his powers. The Roman Senate rejected this proposal, charging him instead to administer (besides Egypt) Spain, Gaul, and Syria for the next 10 years, while it itself supervised the rest of the empire. (It is uncertain whether the Senate named him proconsul for this *provincia* in 27 or expected him to superintend it by virtue of being consul.) Three days later, among other honours, it bestowed upon him the name by which he has ever since been known, Augustus.

Therewith the Roman state ceased to be the plaything of dictators and triumvirs and became a *res publica* again. Indeed, the Latin expression for this settlement of 27 is *res publica restituta,* the usual translation of which, "restoration of the republic," is misleading, because the Latin *res publica* means a state governed by constitutional methods, whether a republic or not.

As most of the troops still under arms were in the regions entrusted to Augustus' charge, the arrangements of 27 BC hardly affected his military strength. Moreover, so long as he was consul (he was reelected every year until 23 BC) he was civilian head of government as well. In other words, he was still preeminent and all-powerful, even if he had, in his own words, placed the *res publica* at the disposal of the Senate and the Roman People. Augustus particularly wished to conciliate the senatorial class, without whose experienced help civilian government was impossible. But his monopolization of the consulship offended the Senate, so that some different arrangement was clearly necessary.

Augustus' powers

Accordingly, in 23 Augustus made a change; he vacated the consulship and never held it again (except momentarily in 5 BC and again in 2 BC, for a limited, specific purpose). In its place he received the tribunician power (*tribunicia potestas*). He could not become an actual plebeian tribune, as Julius Caesar's action in making him a patrician had disqualified him for the office. But he could acquire the rights and privileges pertaining to the office; and they were conferred upon him, apparently by the Senate, whose action was then ratified by the popular assembly. He had already been enjoying some of a tribune's privileges since 36; but he now acquired them all and even some extra ones, such as the right to convene the Senate whenever he chose and to enjoy priority in bringing business before it. Through his tribunician power he could also summon the popular assembly and participate fully in its proceedings. Clearly, although no longer consul, he still retained the legal right to authority in civilian affairs.

The arrangement of 23 entailed an additional advantage. The power of the plebeian tribune was traditionally associated with the protection of citizens, and Augustus' acquisition of it was therefore unlikely to rouse resentment. Indeed, Augustus thenceforth shrewdly propagated the notion that, if his position in the state was exceptional (which it clearly was), then this was precisely because of his tribunician power. Although he held it for only one year at a time, it was indefinitely renewable and was pronounced his for life. Thus, it was both annual and perpetual and was a suitable vehicle for numbering the years of his supremacy. His era (and this is true also of later emperors) was counted officially from the year when he acquired the tribunician power.

The year 23 likewise clarified the legal basis for Augustus' control of his *provincia* (the region under his jurisdiction) and its armed forces. The Senate invested him with an *imperium proconsulare* (governorship and command of an empire); this had a time limit, but it was automatically renewed whenever it lapsed (usually every 10 years). This proconsular *imperium,* furthermore, was pronounced valid inside Italy, even inside Rome and the *pomerium* (the boundary within which only Roman gods could be worshipped and civil magistrates rule), and superior (*majus*) to the *imperium* of any other proconsul. Thus, Augustus could intervene quite legally in any province, even in one entrusted to someone else.

It is clearly this *imperium proconsulare majus* that made Augustus unassailable; but, because it belonged more to military than to civilian life, he preferred not to draw attention to it. In fact, he deemphasized it: it is not reflected in either his official titles or his autobiography, the tribunician power being paraded as the expression of his supreme authority.

Augustus' honours

After 23 no fundamental change in Augustus' position occurred. He felt no need to hold offices that in republican times would have conferred exceptional power (*e.g.,* dictatorship, lifetime censorship, or regular consulship), even though these were offered him. Honours, of course, came his way: in 19 BC he received some consular rights and prerogatives, presumably to ensure that his *imperium* was in no particular inferior to a consul's; in 12, when Lepidus died, he became *pontifex maximus* (he had long since been elected into all of the priestly colleges); in 8 BC, the eighth month of the year was named after him; in 2 BC, he was designated *pater patriae* ("father of his country"), a distinction which he particularly esteemed because it suggested that he was to all Romans what a *paterfamilias* was to his own household. He also accepted special commissions from time to time: *e.g.,* to superintend the supply of grain and water, the maintenance of public buildings (including temples), the regulation of the Tiber, the police and fire-fighting services, and the upkeep of Italy's roads. Such behaviour advertised his concern for public well-being and also the legal correctness of his position; but it hardly altered the fact that his tribunician power and proconsular *imperium,* as defined in 23, were the real bases of his strength and the twin legal supports upon which his position and authority rested. It was these that made him a charismatic leader of unrivalled prestige (*auctoritas*), whose merest suggestions were binding; they ensured absolute imperial predominance for the future.

Like an ordinary Roman, he contented himself with three names. His, however, Imperator Caesar Augustus, were absolutely unique, with a magic all their own that caused all later emperors to appropriate them, at first selectively but after AD 69 in their entirety. Thereby they became titles, reserved for the emperor (or, in the case of the name Caesar, for his heir apparent); from them derive the titles emperor, kaiser, and tsar. Yet, as used by Augustus and his first four successors, the words Imperator Caesar Augustus were names, not titles—that is, respectively, praenomen, nomen (in effect), and cognomen. One title that Augustus did have was *princeps* ("prince"); this, however, was unofficial—a mere popular label, meaning Rome's first citizen—and government documents (inscriptions, coins, etc.) do not apply it to Augustus. But because of it the system of government he devised is called the principate.

The Roman Senate and the urban magistracies. Augustus regarded the Senate, whose leading member (*princeps senatus*) he had become in 28, as a body with important functions; it heard fewer overseas embassies than formerly, but otherwise its dignity and authority seemed unimpaired; its members filled the highest offices; its decrees, although not formally called laws, were just as binding; it soon became a high court, whose verdicts were unappealable; it supervised the older provinces and nominally the state finances as well, and it also in effect elected the urban magistrates; formally, even the emperor's powers derived from the Senate. Nevertheless, it lacked real power. Its provinces contained few troops (and by AD 40 it had ceased to control even these few). Hence, it could hardly dispute Augustus' wishes. Moreover, it was he, in fact, who superintended state finances. Above all, he controlled its membership, every senator's career depending upon his goodwill. But he valued the Senate as the repository of the true Roman spirit and traditions and as the body representing public opinion. He was considerate toward it, shrewdly anticipated its reactions, and generally avoided contention with it. He regularly kept it informed about his activities; and an imperial council (Concilium Principis), which he consulted on matters of policy, in the manner of a republican magistrate seeking the opinion of his advisory committee, consisted of the consuls, certain other magistrates, and 15 senators—not handpicked by him but chosen by lot every six months.

Augustus' relationship with the Senate

To rid the Senate of unworthy elements he reduced its numbers by successive reviews to about 600 (from the triumviral 1,000 or more). Sons of senators and men of good repute and substance who had served in the army and the *vigintiviri* ("board of twenty", minor magistracy) could become members by being elected, at age 25 or over, to the quaestorship. Their subsequent rank in the Senate depended on what other magistracies they managed to win; these were, in ascending order, the aedileship (or plebeian tribunate), the praetorship, and the consulship. No one disliked by Augustus could expect to reach any of them, while anyone whom he nominated or endorsed was sure of election. Despite the Emperor's control, there were enough candidates for keen contests. By AD 5 *destinatio* seems to have been the practice: that is, a special panel of senators and equites selected the praetors and consuls, and the Centuriate Assembly automatically ratified their choice. In *c.* AD 5, likewise, the consulship was shortened to six months. This not only gratified senators and increased the number of high-ranking qualified officials but also showed that the consuls' duties were becoming largely ceremonial. This was also true, but to a far lesser degree, of the other unpaid magistrates. A senator really made his mark in between his magistracies, when he served in important salaried posts, military or civilian or both, sometimes far from Rome.

Rise of equites to a privileged status

The equestrian order. Senators, however, were either too proud or too few to fill all the posts. Some posts were considered menial and went to the Emperor's freedmen or slaves. Others were entrusted to equites, and the equestrian order soon developed into one of the great institutions of the empire. Augustus decided that membership in the order should be open to Roman citizens of means and reputation but not necessarily of good birth. Ultimately,

there were thousands of equites throughout the empire. Although this was a lower aristocracy, a good career was available to them. After tours of duty as an army officer (the so-called *militiae equestres*), an aspiring eques might serve as the emperor's agent (*procurator*) in various capacities and eventually become one of the powerful prefects (of the fleet, of the *vigiles,* or fire brigade, of the grain supply, of Egypt, or of the Praetorian Guard). This kind of an equestrian career became standardized only under Claudius I; but Augustus began the system and, by his use of equites in responsible posts, founded the imperial civil service, which later was staffed chiefly by them. The equites also performed another function: the senatorial order had difficulty in maintaining its numbers from its own ranks and depended on recruitment from below; this meant from the equestrian order, and, as it was not confined to Rome or even to Italy, the Senate gradually acquired a non-Italian element. The western provinces were already supplying senators under Augustus.

Administration of Rome and Italy. Ordinary Roman citizens who were neither senators nor equites were of lesser consequence. Although still used, the old formula Senatus Populusque Romanus had changed its meaning: in effect, its Populusque Romanus portion now meant "the emperor." The "Roman People" had become the "Italian People," and it was embodied in the person of Augustus, himself the native of an Italian town. To reduce the risk of popular demonstrations in Rome, the Emperor provided grain doles, occasional donatives, and various entertainments; but he allowed the populace no real power. After AD 5 the Roman People's participation in public life consisted in the formality of holding occasional assemblies to ratify decisions made elsewhere. Ultimately, this caused the distinction between the Roman citizens of Italy and the provincial inhabitants of the overseas empire to disappear; under Augustus, however, the primacy of Italy was insistently emphasized.

The primacy of Italy

Indeed, Italy and justice for its inhabitants were Augustus' first cares. Arbitrary triumviral legislation was pronounced invalid after 29 BC, and ordinary Roman citizens everywhere had access to Augustus' own court of appeal (his appellate jurisdiction dated from 30 BC and in effect replaced the republican appeal to the People). His praetorian and urban cohorts provided physical security; his officials assured grain supplies; and he himself, with help from such aides as Agrippa, monumentalized Italian towns. The numerous Augustan structures in Italy and Rome (a city of brick before his time and of marble afterward) have mostly perished, but impressive ruins survive (*e.g.,* aqueduct, forum, and mausoleum in Rome; bridge at Narni; arch at Fano; gate at Perugia). Doubtless their construction alleviated unemployment, especially among the proletariat at Rome. But economic considerations did not influence Augustus' policies much (customs tariffs, for instance, were for fiscal, not protective, purposes), nor did he build harbour works at Ostia. Italian commerce and industry nevertheless flourished in the conditions he created. Public finances, mints, and coinage issues, chaotic before him, were placed on a sound basis, partly by the introduction of a sales tax and of a new levy (death duties) on Roman citizens—who hitherto had been subject only to harbour dues and manumission (freeing of slaves) charges—and partly by means of repeated subventions to the public treasury (*aerarium Saturni*) from Augustus' own enormous private resources (*patrimonium Caesaris*). His many highways also contributed to Italy's economic betterment.

Augustus' great achievement in Italy, however, was to restore morale and unify the country. The violence and self-aggrandizement of the 1st century BC had bred apathy and corruption. To reawaken a sense of responsibility, especially in official and administrative circles, Augustus reaffirmed traditional Italian virtues (by outlawing adultery, strengthening family ties, and stimulating the birthrate) and revived ancestral religion (by repairing temples, building new shrines, and reactivating moribund cults and rituals). To infuse fresh blood and energy into disillusioned Roman society, he promoted the assimilation of Italy: the flower of its municipal towns entered the Roman Senate, and Italy became firmly one with Rome. To keep the citizen body pure, he made manumission of slaves difficult, and from those irregularly manumitted he withheld the citizenship.

Administration of the provinces. Sharply distinguished from Italy were the provinces of the empire. From 27 BC on they were of two types. The Senate supervised the long-established ones, the so-called public provinces: their governors were chosen by lot, usually served for a year, commanded no troops, and were called proconsuls (although only those superintending Asia and Africa were in fact former consuls, the others being former praetors). The Emperor supervised all other provinces, and collectively they made up his *provincia:* he appointed their governors, and these served at his pleasure, none with the title of proconsul because in his own *provincia* proconsular *imperium* was wielded by him alone. These imperial provinces might be "unarmed," but many of them were garrisoned, some quite heavily. Those containing more than one legion were entrusted to former consuls and those with a legion or less to former praetors; in both cases their governors were called *legati Augusti pro praetore* ("legates of Augustus with authority of a praetor"). There were also some imperial provinces governed not by senators but by equites (usually styled procurators but sometimes prefects); Judaea at the time of Christ's crucifixion was such an equestrian province, Pontius Pilate being its governor. An entirely exceptional imperial province was Egypt, so jealously guarded that no senator could visit it without express permission; its prefect was unique in being an equestrian in command of legions.

Two types of provinces

The provinces paid tribute, which helped to pay for the armed services, the growing civil service, and the public-works programs. Periodical censuses, carefully listing provincial resources, provided the basis for the two direct taxes: *tributum soli,* exacted from occupiers of provincial soil, and *tributum capitis,* paid on other forms of property (it was not a poll tax, except in Egypt and in certain backward areas). In addition, the provinces paid indirect taxes, such as harbour dues. In imperial provinces the direct taxes (*tributa*) were paid to the Emperor's procurator, an equestrian official largely independent of the governor. In senatorial provinces, quaestors supervised the finances; but, increasingly, imperial procurators also appeared. The indirect taxes (*vectigalia*) were still collected by *publicani,* but they were now much more rigorously controlled and were being gradually replaced by imperial civil servants.

As under the republic, each province was a collection of autonomous communities (*civitates*), roughly resembling modern counties, and it was through them that Rome governed its empire; without them provincial administration would have simply broken down. The autonomous communities, apart from a privileged few, all paid taxes to Rome. Most of them were peregrine; that is, their burgesses were not Roman citizens; but these non-Roman communities tended to decrease in number because they gradually adopted Rome's language and ways and thereby became qualified to receive "Latin rights" (*jus Latium,* or *jus Latii*)—a status that prepared them for the eventual bestowal of full Roman citizenship. Peregrine communities upgraded in this way were called *municipia.*

Autonomous communities

Whole communities with Roman citizenship in the provinces were virtually a new development, because there had been hardly any before Julius Caesar's dictatorship. Those called *coloniae* (see below) ranked highest, being regarded as actual extensions of Rome itself: some of them were even exempted from payment of tribute. There were, however, few *coloniae* and practically no *municipia* in the East.

In both senatorial and imperial provinces the *civitates* enjoyed local self-government. The Roman and Latin communities were city-states, resembling the towns of Italy; they regularly had an assembly of adult males and a local council (*ordo*), and their chief officials—collegial, annual, and unpaid—were either *duoviri* (in *coloniae*) or *quattuorviri* (in *municipia*). The peregrine communities might or might not be city-states—many were organized on tribal, not municipal, lines; moreover, the city-states among them showed great variety, some of the Greek-

speaking ones having constitutions that were centuries old. Although Rome regularly ensured control by the propertied classes, it rarely imposed Roman-style constitutions. It did perhaps favour urbanized communities, rural areas being subordinated to city-states for administrative purposes until ready for municipal institutions of their own. Naturally, too, peregrine communities tended to imitate nearby Roman towns, so that even some tribally organized *civitates* boasted *duoviri* and an *ordo.*

The provinces were generally better off under the empire. The governing power now felt some responsibility for the governed and sent out salaried and experienced officials. These were closely supervised. During the first two centuries, provincial communities flourished on the whole; in many, a sense of civic liberty and healthy local patriotism induced wealthy burgesses to build baths, theatres, and the like at their own expense. Above all, there was peace, the Pax Romana.

Emperor worship. For this priceless gift of peace, many individuals and even whole communities, in Italy and elsewhere, expressed their thanks spontaneously by worshipping Augustus and his family. Emperor worship was also encouraged officially, however, as a focus of common loyalty for the polyglot empire. In the provinces, to emphasize the superiority of Italy, the official cult was dedicated to Roma et Augustus; to celebrate it, representatives from provincial communities or groups of communities met in an assembly (Concilium Provinciae), which incidentally might air grievances as well as satisfactions. (This system began in the Greek-speaking provinces, long used to wooing their rulers with divine honours. It penetrated the west only slowly; but from 12 BC an assembly for the three imperial Gallic provinces existed at Lugdunum.) In Italy the official cult was to the Genius Augusti (the life spirit of his family); it was coupled in Rome with the Lares Compitales (the spirits of his ancestors). Its principal custodians (*seviri Augustales*) were normally freedmen. Calculation always entered into emperor worship; it was no act of piety but a political gesture, which could be continued after the emperor's death if the Senate pronounced him worthy of posthumous deification.

The army. It was Augustus' soldiers, however, not his worshippers, that made him all-powerful. Their allegiance, like the name Caesar, was inherited from his "father," the deified Julius. The allegiance was to the Emperor personally, through a military oath taken in his name every January 1; and the soldiers owed it after his death to his son or chosen successor. This preference of theirs for legitimacy could not be ignored, because they were now a standing army, something that the republic had lacked. Demobilization reduced the 60 legions of Actium to 28, a number hardly sufficient but all that Augustus' prudence or economy would countenance. These became permanent formations, each with its own number and name; the soldiers serving in them were called legionaries. Besides the legionaries there were auxiliaries (or supporting troops), approximately as numerous. The two corps together numbered about 300,000 men. To them must be added the garrison of Italy—the praetorian cohorts, or emperor's bodyguard, about 10,000 strong—and the marines of the imperial fleet, which had its main headquarters at Misenum and Ravenna in Italy and subsidiary stations and flotillas on seas and rivers elsewhere (the marines, however, were not reckoned good combat forces). All these troops were long-service professionals—the praetorians serving 16 years; legionaries, 20; auxiliaries, 25; and marines, 28—with differing pay scales, the praetorians' being highest. In addition to their pay, the men received donatives, shares of booty, and retirement bonuses from a special treasury (*aerarium militare*) established in AD 6 and maintained out of the sales tax and Roman citizens' death duties. Under Augustus the praetorians were normally Italians, but many legionaries and virtually all auxiliaries were provincials, mainly from the imperial provinces in the west, the legionaries coming from municipal towns and the auxiliaries from tribal areas. The tendency to use provincials grew, and by the year 100 the Roman imperial army was overwhelmingly non-Italian. Nevertheless, it helped greatly to Romanize the empire. The legionaries were Roman citizens from the day they enlisted, if not before, and the auxiliaries (after Claudius anyway) from the day they were discharged; and, though serving soldiers could not legally marry, many had mistresses whose children often became Roman citizens. The troops, other than praetorians and marines, passed their years of service in the "armed" imperial provinces—the auxiliaries in forts near the frontier and the legionaries at some distance from it in camps that showed an increasing tendency, especially after AD 69, to become permanent (some of them, indeed, developed into great European cities). There was no central reserve, because, although desirable for emergencies, it might prove dangerous in peacetime.

Allegiance of the army to the Emperor

The officers were naturally Roman citizens. In the legions those of the highest rank (*legati* and *tribuni*) were senators or equites; lower officers (*centuriones*) might enter directly from Italian or provincial municipalities or might rise through the ranks, and by the time they retired, if not sooner, many of them were equites. In the auxiliaries the unit commanders (*praefecti*) were equites, often of provincial birth. On retirement the soldiers frequently settled in the provinces where they had served, made friends, and perhaps acquired families; imperial policy favoured this practice. Indeed, many emperors founded colonies of veterans in the provinces, establishing the men in Roman communities of regular municipal type with free grants of land. Some of Augustus' *coloniae* were in Italy; but many more were overseas, as those of his successors regularly were. These military *coloniae* were the most highly regarded type of urban community under the empire and excellent Romanizing agents.

Foreign policy. After Actium and on two other occasions, Augustus solemnly closed the gates of the shrine of Janus (a gesture of peace), to show that Rome had peace as well as a *princeps.* These well-publicized gestures were purely temporary; the gates were swiftly reopened. His proconsular *imperium* made Augustus the arbiter of peace and war, and an ostensible search for defensible frontiers made his a very warlike reign. He sought to repair the omission of the republic that had left the limits of Roman territorial claims rather vague and indefinite; he planned conquests stretching to the boundaries defined by nature (deserts, rivers, and ocean shores), not always, however, with immediate annexation in mind. When annexation did occur, it was followed by the construction of solidly built military roads, paved with thick stone blocks: these also served the official post system (*cursus publicus*) and were provided with rest stages and overnight lodges at regular intervals.

Territorial expansion under Augustus

Areas where subjugation looked arduous and where Romanization seemed problematic were left to client kings, dependent on the Emperor's support and goodwill and under obligation to render military aid to Rome. Such satellite kingdoms spared Augustus the trouble and expense of maintaining strong defenses everywhere; nevertheless, their ultimate and intended destiny was incorporation as soon as it suited their overlord's convenience. Usually, territory was gained more easily by creating and subsequently incorporating a client kingdom than by launching an expansionist war.

In the south, Augustus found suitable frontiers quickly. In 25 BC an expedition under Aelius Gallus opened the Red Sea to Roman use and simultaneously revealed the Arabian Desert as an unsurpassed and, indeed, unsurpassable boundary. The same year Gaius Petronius, the prefect of Egypt, tightened Rome's grip as far as the First Cataract and established a broad military zone beyond it. The vast region north of the Sahara and the Atlas Mountains was also secured (*c.* 25) after a series of punitive raids against native tribes and the annexation of one client kingdom (Numidia) and the creation of another (Mauretania). Three legions, two in Egypt and one in Africa (a senatorial province), policed the southern shore of the Mediterranean.

In the west, consolidation was extended to the Atlantic. Gaul, Julius Caesar's conquest, was organized as four provinces: senatorial Narbonensis and the imperial three Gauls (Aquitania, Belgica, and Lugdunensis). In Spain, after Agrippa successfully ended in 19 BC the terminal

campaign that Augustus had launched in person in 26, three provinces were formed: senatorial Baetica and imperial Lusitania and Tarraconensis. Three legions enforced Roman authority from Gibraltar to the mouth of the Rhine. Augustus ignored the advice of court poets and others to advance still farther and annex Britain.

Difficulties in the east and north

In the east, Parthia had demonstrated its power against Crassus and Antony, and Augustus proceeded warily. He retained Antony's ring of buffer client kingdoms, although he incorporated some, including the most celebrated of them, Judaea; he made Judaea a province in AD 6, respecting some of the customs of its Jewish inhabitants in the process. Augustus stationed four legions in Syria and obviously envisaged the Euphrates River and the northern extension of the Arabian Desert as the desirable frontier with Mesopotamia. Farther north, however, no such natural line existed. North of the Black Sea the client kingdom of the Cimmerian Bosporus, under its successive rulers Asander and Polemo, helped to contain southward and westward thrusts by the Scythians, an Iranian people related to the Parthians, and this provided protection in the north for Asia Minor and its provinces (senatorial Asia and Bithynia-Pontus and imperial Cilicia and Galatia, the latter a large new province created in 25 BC out of Amyntas' client kingdom). By a show of force, Augustus' stepson Tiberius in 20 recovered the standards lost at Carrhae and installed Tigranes as client king of Armenia, and Augustan propaganda depicted this as a famous victory; but strategic considerations inevitably obliged the Parthians, once they settled their internal, dynastic dissensions, to dispute Roman control of Armenia, so that Augustus hardly settled the eastern frontier. Missions were sent to the East repeatedly (Agrippa, 17–13 BC; Gaius Caesar, AD 1–4; Germanicus, 18–19), and Armenia remained a problem for Augustus' successors: Tiberius successfully maintained Roman influence there, but Gaius and Claudius failed to do so, leaving Nero with a difficult situation.

In the north, too, there was difficulty. The Alps and their passes were finally subjugated early in Augustus' reign. This enabled Tiberius and his brother Drusus between 16 and 8 BC to conquer all the way to the great rivers of central Europe. New provinces were created in the Alps and Tyrol (Maritime and Pennine Alps, Raetia, Noricum) and also farther east (Pannonia, Moesia). Stability along the Danube was precariously maintained, under Augustus and later, by means of periodical alliances with Maroboduus and his successors, who ruled Germanic tribes like the Marcomanni and Quadi in Bohemia to the north of the river, and by the existence of a Thracian client kingdom to the south of its lowest course. The push across the Rhine began in 12 BC; it reached the Elbe, but consolidation beyond the Rhine proved elusive. A revolt in Pannonia (AD 6–9) interrupted it, and, in AD 9, German tribes under Arminius annihilated Quinctilius Varus and three legions in the Teutoburg Forest. This disaster reduced the number of legions to 25 (it did not reach 28 again until half a century later), and it disheartened Augustus. Old and weary, he withdrew to the Rhine and decided against all further expansion, a policy he urged upon his successor. For the watch on the Rhine the military districts of Upper and Lower Germany were created, containing eight legions between them. Another seven garrisoned the Danubian provinces. These figures reveal imperial anxiety for the northern frontier.

Effects of peace

Economic life. Although widespread, Augustus' wars chiefly affected the frontier districts. Elsewhere, peace prevailed. Indeed, never before had so large an area been free of war for so long. This state of affairs helped trade. The suppression of piracy and the use of military roads, which the frontier warfare itself brought into being, provided safe arteries of commerce. Stable currency also aided economic growth. Activity directly connected with the soil predominated; but there were also many establishments, usually small, engaged in manufacturing, and such products as textiles, pottery, tiles, and papyrus were turned out in surprising quantities. Advanced techniques were also known: glassblowing, for example, dates from the Augustan age. Most products were consumed locally, but the specialties or monopolies from any region usually exceeded its needs, and the surplus was sold elsewhere, generating a brisk interchange of goods. Some travelled great distances, even beyond the empire: trade with India, for example, reached respectable proportions once the nature of the monsoon was understood, and the Red Sea was opened to Roman shipping. Merchants, especially Levantines, travelled everywhere, and fairs were frequent. The Mediterranean world was linked together as never before, and standardization made considerable headway. In Augustus' day Italy was economically the most important part of the empire. It could afford to import on a large scale, thanks partly to provincial tribute but above all to its own large productivity. The eastern provinces, for their part, recovered rapidly from the depredations of the civil wars and were industrially quite advanced. The other provinces were less developed, but they soon ceased being mere suppliers of raw materials; they learned to exploit their natural resources by using new techniques and then began overtaking the more advanced economies of Italy and the Greek-speaking regions. The importance of trade in unifying the empire should not be underestimated.

Augustan art and literature. In 17 BC Rome held Secular Games, a traditional celebration to announce the entry into a new epoch (*saeculum*). New it was, for, though Augustus preserved what he could of republican institutions, he added much that was his own. His Rome had become very Italian, and this spirit is reflected in the art and literature of his reign. Its greatest writers were native Italians, and, like the ruler whose program they glorified, they used the traditional as the basis for something new. Virgil, Horace, and Livy imitated the writing of classical Greece, but chiefly in outward form, their tone and outlook being unHellenic. The glory of Italy and faith in Rome are what inspired Virgil's *Georgics* and *Aeneid,* Horace's *Odes,* and the first 10 books of Livy's history.

In Augustan art a similar fusion was achieved between the prevailing Attic and Hellenistic models and Italian naturalism. The sculptured portraits on the Ara Pacis Augustae of 9 BC, for all their lifelike quality, are yet in harmony with the classical poise of the figures, and they strike a fresh note: the stately converging processions (Rome's imperial family and magistrates on one side; senators, equites, and citizens on the other) became the prototypes for all later processional reliefs. Augustan painting likewise displays a successful combination of Greek and Roman elements, to judge from the frescoes in the house of Livia on the Palatine. In Augustan architecture Greek features abound, elaborated on basically Italic plans. Unless Agrippa's Pantheon was domed, there is not as yet significant evidence of the soaring vaults and domes that Roman concrete could have already made possible. Building was, however, very active and widespread.

Cultural excellence of the Augustan age

The culture of the age undoubtedly attained a high level of excellence, dominated by the personality of the Emperor and his accomplishments. Imperial art had already reached full development, a matter of no small moment, because Rome's political predominance made the spread of its influence inevitable. The Mediterranean world was soon assuming a Roman aspect, and this is a measure of Augustus' extraordinary achievement. Yet it was an achievement with limitations. His professed aim—to promote stability, peace, security, and prosperity—was irreproachable; but perhaps it was also unexciting. Emphasizing conservatism by precept and his own example, he encouraged the simpler virtues of a less sophisticated age, and his success made this sedate but rather static outlook fashionable. Men accepted the routine of his continuing rule, at the cost, however, of some loss of intellectual energy and moral fervour. The great literature, significantly, belongs to the years near Actium, when men's imagination still nursed heady visions of Roman victory and Italian destiny. After the Secular Games the atmosphere became more commonplace and produced the frivolities of Ovid and the pedestrian later books of Livy.

Appraisal of Augustus. Augustus' position as *princeps* cannot be defined simply. He was neither a Roman king (*rex*) nor a Hellenistic monarch (*basileus*), nor was he, as the 19th-century German historian Theodor Mommsen thought, a partner with the Senate in a dyarchy. He posed

as the first servant of an empire over which the Roman Senate presided, and it would appear that his claim to have accepted no office inconsistent with ancestral custom was literally true. Proconsular *imperium* was a republican institution; and, although tribunician power was not, it contained nothing specifically unrepublican. But, while precedents can be cited for Augustus' various powers, their concentration and tenure were absolutely unparalleled. Under the republic, powers like his would have been distributed among several holders, each serving for a limited period with a colleague. Augustus wielded them all, by himself, simultaneously and without any time limit (in practice, at least). This fact made him an emperor, but it did not necessarily make him a military tyrant.

In discharging both military and civilian functions, Augustus was no different from republican consuls or praetors. Admittedly his military power was overwhelming; but, if he chose not to brandish it, the tone of his reign could remain essentially civilian. Constitutional safeguards were indeed lacking; everything was at the emperor's discretion, and even Augustus passed legislation that made anti-imperial behaviour, real or suspected, treasonable (men were, in fact, executed for conspiracy during his reign). But there had been no constitutional safeguards in the republic, under Sulla, Pompey, the triumvirs, or even Julius Caesar. Augustus' improved police services probably made lower class Romans at least feel safer under him. Augustus had not even deprived the Romans of "liberty," if *libertas* be understood as meaning freedom from arbitrary rule and arbitrary interference.

The principate was something personal, what the emperor chose to make it, and the relations prevailing between emperor and Senate usually indicated what a reign was like. In Augustus' case they reveal a regime that was outwardly constitutional, generally moderate, and certainly effective. But, as he himself implied at the end of his life, he was a skillful actor in life's comedy. Later emperors might have lacked his sureness of touch.

Augustus' accomplishments

When Augustus died, the Senate unhesitatingly pronounced him *divus*—the deified one who had restored peace, organized a standing army to defend the frontiers, expanded those frontiers farther than any previous Roman, improved administrative practices everywhere, promoted better standards of public and private behaviour, integrated Rome and Italy, embellished Rome, reconciled the provinces, expedited Romanization, and above all maintained law and order while respecting republican traditions.

Augustus' luck was hardly inferior to his statecraft. Despite indifferent health, he headed the Roman state in one capacity or another for 56 years. His rule, one of the longest in European history, consolidated the principate so firmly that what might have been an episode became an epoch. At his death there was practically no one left with any personal memory of the republic, and Augustus' wish came true: he had fashioned a lasting as well as constitutional government. The principate endured with only minor changes for about 200 years.

The other Julio-Claudian emperors

The succession. Augustus' first four successors were all related to him by either birth or marriage; he and they collectively formed the Julio-Claudian dynasty (27 BC–AD 68). There being no constitutional provision for the succession, none of them reached the purple by legal right. The Emperor's position was personal, and, on his death, revival of the republic was in theory possible; in practice, however, it was out of the question—even in Augustus' lifetime, appointment of a senator as city prefect was necessary, to deputize for the Emperor during his absences from Rome. Consequently, Augustus began thinking early about who should follow him. The soldiers' views on legitimacy reinforced his own natural desire to found a dynasty; but he had no son and was therefore obliged to select his successor. Death played havoc with his attempts to do so. His nephew Marcellus, his son-in-law Agrippa, his grandsons Gaius and Lucius (Julia's children by Agrippa), were groomed in turn; but they all predeceased him. Augustus, finally and reluctantly, chose a member of the republican nobility, his stepson Tiberius, a scion of the ultra-aristocratic Claudii. In AD 4 Augustus adopted Tiberius as his son and had tribunician power and probably proconsular *imperium* as well conferred upon him. This arrangement was confirmed in 13, and when Augustus died the following year, Tiberius automatically became emperor.

Tiberius (ruled 14–37), during whose reign Christ was crucified, was a soldier and administrator of proved capacity but of a reserved and moody temperament that engendered misunderstanding and unpopularity. Slander blamed him for the death in 19 of his nephew and heir apparent, the popular Germanicus; and, when informers (*delatores*), who functioned at Rome like public prosecutors, charged notables with treason, Tiberius was thought to encourage them. By concentrating the praetorian cohorts in a camp adjoining Rome he increased the soldiers' scope for mischief-making without finding any real security, and in 26 he left Rome permanently for the island of Capreae (Capri), entrusting Rome to the care of the city prefect. Tiberius heeded the aged Augustus' advice and did not extend the empire. (The annexation of Cappadocia, a client kingdom, represented no departure from Augustan policy.) In general he took his duties seriously; but by administering the empire from Capreae he offended the Senate and was never fully trusted, much less really liked. At his death he was not pronounced *divus*. His great-nephew, Germanicus' son Gaius, succeeded him.

Gaius (better known by his nickname, Caligula, meaning Little Boot) ruled from 37 to 41 with the absolutism of an Oriental monarch: his short reign was featured by reckless spending, callous murders, and humiliation of the Senate. Gaius' foreign policy was inept. Projected annexation proved abortive in Britain; it touched off heavy fighting in Mauretania. In Judaea and Alexandria, Gaius' contemptuous disregard of Jewish sentiment provoked near rebellion. When assassination ended his tyranny, the Senate contemplated restoration of the republic but was obliged by the Praetorian Guard to recognize Claudius, Germanicus' brother and therefore Gaius' uncle, as emperor.

Centralization and expansion under Claudius I

Claudius I (ruled 41–54) went far beyond Augustus and Tiberius in centralizing government administration and, particularly, state finances in the imperial household. His freedmen secretaries consequently acquired great power; they were in effect directors of government bureaus. Claudius himself displayed much interest in the empire overseas; he enlarged it significantly, incorporating client kingdoms (Mauretania in 42; Lycia, 43; Thrace, 46) and, more important, annexing Britain. Conquest of Britain began in 43, Claudius himself participating in the campaign; the southeast was soon overrun, a *colonia* established at Camulodunum (Colchester) and a *municipium* at Verulamium (St. Albans), while Londinium (London) burgeoned into an important entrepôt. Claudius also promoted Romanization, especially in the western provinces, by liberally granting Roman citizenship, by founding *coloniae,* and by inducting provincials directly into the Senate—he became censor in 47 and added the senators he wanted, bestowing appropriate quaestorian or praetorian rank upon them to spare the maturer ones among them the necessity of holding junior magistracies; lest existing senators take offense, he elevated some of them to patrician status (a form of patronage often used by later emperors). Claudius' provincial policies made the primacy of Italy less pronounced, although that was hardly his aim. In fact, he did much for Italy, improving its harbours, roads, and municipal administration and draining its marshy districts. The execution of many senators and equites, the insolence and venality of his freedmen, the excessive influence of his wives, and even his bodily infirmities combined to make him unpopular. Nevertheless, when he died (murdered probably by his fourth wife, Agrippina, Augustus' great-granddaughter, who was impatient for the succession of the 16-year-old Nero, her son by an earlier marriage), he was pronounced *divus*.

Nero (ruled 54–68) left administration to capable advisers for a few years but then asserted himself as a vicious despot. He murdered successively his stepbrother Britannicus, his mother Agrippina, his wife Octavia, and his tutor Seneca. He also executed many Christians, accusing them of starting the great fire of Rome in 64 (this

is the first recorded Christian persecution). In Rome his reliance on Oriental favourites and his general misgovernment led to a conspiracy by Gaius Calpurnius Piso in 65, but it was suppressed, leading to yet more executions; the victims included the poet Lucan. The empire was not enlarged under this unwarlike emperor, but it was called upon to put down serious disorders. In Britain in 60–61 the rapacity and brutality of Roman officials provoked a furious uprising under Queen Boudicca; thousands were slaughtered, and Camulodunum, Vernulamium, and Londinium were destroyed. In the east, a major military effort under Corbulo, who was Rome's foremost general, was required (62–65) to reestablish Roman prestige; a compromise settlement was reached with the Romans accepting the Parthian nominee in Armenia and the Parthians recognizing him as Rome's client king. In 66, however, revolt flared in Judaea, fired by Roman cruelty and stupidity, Jewish fanaticism, and communal hatreds; the Prefect of Egypt, Julius Alexander, prevented involvement of the Jews of the Diaspora. A strong army was sent to Judaea under Titus Flavius Vespasianus to restore order; but it had not completed its task when two provincial governors in the west rebelled against Nero—Julius Vindex in Gallia Lugdunensis and Sulpicius Galba in Hispania Tarraconensis. When the praetorians in Rome also renounced their allegiance, Nero lost his nerve and committed suicide. He brought the Julio-Claudian dynasty to an ignominious end by being the first emperor to suffer *damnatio memoriae*—his reign was officially stricken from the record by order of the Senate.

The end of the Julio-Claudian dynasty

Growth of the empire under the Flavians and Antonines. *The year of the four emperors.* Nero's death ushered in the so-called year of the four emperors. The extinction of the Julio-Claudian imperial house robbed the soldiers of a focus for their allegiance, and civil war between the different armies ensued. The army of Upper Germany, after crushing Vindex, urged its commander, Verginius Rufus, to seize the purple for himself. But he elected to support Galba—scion of a republican patrician family claiming descent from Jupiter and Pasiphae—who was recognized as emperor by the Senate. A treasury emptied by Nero's extravagance entailed economy, and this bred unpopularity for Galba; his age (73) was also against him, and unrest grew. Early in January 69 the Rhineland armies acclaimed Aulus Vitellius, commander in Lower Germany; at Rome the praetorians preferred Marcus Salvius Otho, whom Galba had alienated by choosing a descendant of the old republican aristocracy for his successor. Otho promptly procured Galba's murder and obtained senatorial recognition; this ended the monopoly of the purple that the republican nobility had hitherto enjoyed.

Otho, however, lasted only three months; defeated at Bedriacum, near Cremona in northern Italy, by Vitellius' powerful Rhineland army, he committed suicide (April 69). The Senate thereupon recognized Vitellius; but the soldiers along the Danube and in the east supported Vespasianus, the commander in Judaea. In a second battle near Bedriacum, Vitellius' troops were defeated in their turn, and on his death soon afterward an accommodating Senate pronounced Vespasian emperor.

The Flavian emperors. On December 22, 69, the Senate conferred all the imperial powers upon Vespasian *en bloc* with the famous Lex de Imperio Vespasiani, and the Assembly ratified the Senate's action. This apparently was the first time that such a law was passed; a fragmentary copy of it is preserved on the Capitol in Rome.

Vespasian (ruled 69–79) did not originate from Rome or its aristocracy. His family came from Sabine Reate, and with his elevation the Italian bourgeoisie came into its own. He and his two sons, both of whom in turn succeeded him, constituted the Flavian dynasty (69–96). Vespasian faced the same difficult task as Augustus—the restoration of peace and stability. The disorders of 69 had taken troops away from the Rhine and Danube frontiers. Thereupon, the Danubian lands were raided by Sarmatians, a combination of tribes who had overwhelmed and replaced the Scythians, their distant kinsmen, in eastern Europe. The assailants were repelled without undue difficulty, but, because the region between the Rivers Tisa and Danube was now firmly in the hands of the Sarmatian Iazyges, they were an omen for the future.

Developments in the Rhineland were more immediately serious. There in 69 a certain Civilis incited the Batavians serving as auxiliaries in the Roman army to rebel. Gallic tribes joined the movement, and the insurgents boldly overran all but two of the legionary camps along the Rhine. Vespasian sent his relative Petilius Cerealis to deal with the rebels, who, fortunately for Rome, were not united in their aims; and by 70 Cerealis had restored order. That same year Vespasian's elder son, Titus, brought the bloody war in Judaea to its end by besieging, capturing, and destroying Jerusalem.

To rehabilitate the public finances, Vespasian introduced new imposts, including a poll tax on Jews, and practiced stringent economies. With the Senate he was courteous but firm. He allowed it little initiative but used it as a reservoir from which to obtain capable administrators. To that end he assumed the censorship and added senators on a larger scale than Claudius had done, especially from the municipalities of Italy and the western provinces. Already before 69 an aristocracy of service had been born, and the provincialization of the Roman Senate had begun; thereafter this development made rapid headway. Besides the censorship, Vespasian also often held the consulship, usually with Titus as his colleague. His object presumably was to ensure that his own parvenu Flavian house outrank any other. In this he succeeded; the troops especially were ready to accept the Flavians as the new imperial family. On Vespasian's death in 79, Titus, long groomed for the succession, became emperor and immediately had his father deified.

Vespasian's policies

Titus (ruled 79–81) had a brief reign, marred by disasters (the volcanic eruption that buried Pompeii and Herculaneum and another great fire in Rome); but his attempts to alleviate the suffering and his general openhandedness won him such popularity that he was unhesitatingly deified after his early death.

Domitian (ruled 81–96), Titus' younger brother, had never been formally indicated for the succession; but the praetorians acclaimed him, and the Senate ratified their choice. Throughout his reign Domitian aimed at administrative efficiency, but his methods were very high-handed. For him the Senate existed merely to supply imperial servants. He also used equites extensively, more than any previous emperor. He held the consulship repeatedly, was *censor perpetuus* from 85 on, and demanded other extravagant honours. On the whole, his efficiency promoted the welfare of the empire. Above all, he retained the allegiance of the troops. Although scornful of the Senate's dignity, he insisted on his own and mercilessly punished any act of disrespect, real or fancied, toward himself. He became even more suspicious and ruthless when Saturninus, commander in Upper Germany, attempted rebellion in 89. He crushed Saturninus; executions and confiscations ensued, and *delatores* flourished. The tyranny was particularly dangerous to senators, and it ended only with Domitian's assassination in 96. The Flavian dynasty, like the Julio-Claudian, ended with an emperor whose memory was officially damned.

The disorders in 69 were the cause of some military reforms. Under the Flavians, auxiliaries usually served far from their native heaths under officers of different nationality from themselves. At the same time, the tasks assigned to them came increasingly to resemble those performed by the legionaries. The latter grew less mobile, as stone camp buildings came to be the rule; and it became common for detachments from a legion (*vexillationes*), rather than the entire legion, to be used for field operations. This new-type army proved its mettle in Britain, where the advance halted by Boudicca's revolt was now resumed. Between 71 and 84 three able governors—Petilius Cerealis, Julius Frontinus, and Julius Agricola, the latter Tacitus' father-in-law—enlarged the province to include Wales and northern England; Agricola even reached the Scottish highlands before Domitian recalled him.

Military reforms under the Flavians

Along the Rhine, weaknesses revealed by Civilis' revolt were repaired. Vespasian crossed the river in 74 and annexed the Agri Decumates, the triangle of land between

the sources of the Rhine and the Danube. To consolidate the position, he and Domitian after him penetrated the Neckar Valley and Taunus Mountains, and fortifications began to take shape to the east of the Rhine, a military boundary complete with strongpoints, watchtowers, and, later, a continuous rampart of earthworks and palisades. Once Saturninus' revolt in 89 had been suppressed, Domitian felt the situation along the Rhine sufficiently stable to warrant conversion of the military districts of Upper and Lower Germany into regular provinces and the transfer of some Rhineland troops to the Danube. To the north of this latter river, the Dacians had been organized into a strong kingdom, ruled by Decebalus and centring on modern Romania; in 85 they raided southward across the Danube, and in the next year they defeated the Roman punitive expedition. Domitian restored the situation in 88, but Saturninus' rebellion prevented him from following up his success. Domitian and Decebalus thereupon came to terms: Decebalus was to protect the lower Danube against Sarmatian attack, and Domitian was to pay him an annual subsidy in recompense. The Danubian frontier, however, remained disturbed, and Domitian wisely strengthened its garrisons; by the end of his reign it contained nine legions, as against the Rhineland's six, and Pannonia was soon to become the military centre of gravity of the empire.

Strengthening of the eastern frontier

The Flavians also took measures to strengthen the eastern frontier. In Asia Minor, Vespasian created a large "armed" province by amalgamating Cappadocia, Lesser Armenia, and Galatia; and the whole area was provided with a network of military roads. South of Asia Minor, Judaea was converted into an "armed" province by getting legionary troops; and two client kingdoms—Commagene and Transjordan—were annexed and added to Syria. Furthermore, the legionary camps seem now to have been established right on the Euphrates at the principal river crossings. This display of military strength kept the empire and Parthia at peace for many years.

The Antonine emperors. Marcus Cocceius Nerva, an elderly senator of some distinction, was the choice of Domitian's assassins for emperor; and the Senate promptly recognized him. The soldiers did so much more reluctantly, however; and, because the year 69 had revealed that emperors no longer needed to be Roman aristocrats and could be chosen elsewhere than at Rome, their attitude imposed caution.

Nerva, who ruled from 96 to 98, adopted a generally lavish and liberal policy, but it failed to win the soldiers over completely, and he proved unable to save all Domitian's murderers from their vengeance. Unrest subsided only when, overlooking kinsmen of his own, he adopted an outstanding soldier, Marcus Ulpius Trajanus, who was governor of Upper Germany, as his successor. Nerva himself died a few months later.

Trajan (ruled 98–117) was the first and perhaps the only emperor to be adopted by a predecessor totally unrelated to him by either birth or marriage. He was also the first in a series of "good" rulers who succeeded one another by adoption and for most of the 2nd century provided the empire with internal harmony and careful government; they are collectively, if somewhat loosely, called the Antonine emperors. More significantly still, Trajan, a Spaniard, was also the first *princeps* to come from the provinces; with the greater number of provincials now in the Senate, the elevation of one of them, sooner or later, was practically inevitable. Throughout his reign, Trajan generally observed constitutional practices. Mindful of the susceptibilities of the Senate, he regularly consulted and reported to it. Modest in his bearing, he did not claim ostentatious honours such as frequent consulships or numerous imperial salutations, and he mixed easily with senators on terms of cordial friendship. This reestablished mutual respect between *princeps* and Senate. Empire and liberty, in Tacitus' words, were reconciled, and the atmosphere of suspicion, intrigue, and terror surrounding the court in Domitian's day disappeared. Trajan endeared himself also to the populace at large with lavish building programs, gladiatorial games, and public distributions of money. Above all, he was popular with the armed forces; he was the soldier-emperor *par excellence.* Understandably, he received the title Optimus (Best), officially from 114 on (and unofficially for many years earlier).

Yet Trajan was a thoroughgoing autocrat who intervened without hesitation or scruple even in the senatorial sphere, whenever it seemed necessary. His aim was efficiency; his desire was to promote public welfare everywhere. He embellished Rome with splendid and substantial structures, and he showed his care for Italy by refurbishing and enlarging the harbours at Ostia, Centumcellae, and Ancona. He sent officials called *curatores* to Italian municipalities in financial difficulties and helped to rehabilitate them. He greatly expanded an ingenious charity scheme, which Nerva had probably begun: money was loaned to farmers on easy terms, and the low interest they paid went into a special fund for supporting indigent children. Nor did Trajan neglect Italy's highway network: he built a new road (Via Traiana) that soon replaced the Via Appia as the main thoroughfare between Beneventum and Brundisium.

Interest in Italy implied no neglect of the provinces. *Curatores* were also sent to them; and to rescue Achaea and Bithynia, senatorial provinces, from threatened bankruptcy, Trajan made them both temporarily imperial, sending special commissioners of his own to them. His correspondence with his appointee in Bithynia, the younger Pliny, has survived and reveals how meticulously the Emperor superintended even the smallest details.

Trajan's sense of responsibility cannot be doubted, and his well-meaning example inspired his immediate successors to imitate his paternalism. Such imperial encroachment, however, helped to create a huge and burdensome bureaucracy and to stifle the local initiative on which municipal autonomy throughout the empire depended.

Expansion of the empire under Trajan

Trajan's efforts were as notable in the military sphere as in the civilian. He decided to strengthen the dangerous Danube frontier by converting Dacia into a salient of Roman territory north of the river in order to dismember the Sarmatian tribes and remove the risk of large, hostile combinations to a safer distance. Accordingly, he mustered a large army, conquered Decebalus in two hard-fought wars (101–102; 105–106), and annexed Dacia and filled it with settlers from neighbouring parts of the empire. On the eastern frontier he planned a similar operation, evidently in the conviction, shared by many eminent Romans both before and after him, that only conquest could solve the Parthian problem. Possibly, too, he wished to contain the menace of the Sarmatian Alani in the Caspian region. In a preliminary move, the Nabataean kingdom of Arabia Petraea was annexed in 105–106. Then in 114 Trajan assembled another large army, incorporated the client kingdom of Armenia, and invaded Parthia.

After spectacular victories in 115 and 116, he created additional provinces (Northern Mesopotamia, Assyria) and reached the Persian Gulf. But he had merely overrun Mesopotamia; he had not consolidated it, and, as his army passed, revolts broke out in its rear. The Jews of the Diaspora and others seized their chance to rebel, and before the end of 116 much of the Middle East besides Parthia was in arms (Cyrene, Egypt, Cyprus, Asia Minor). Trajan proceeded resolutely to restore the situation, but death found him still in the East.

Before his last illness, he had not formally indicated his successor. But high honours and important posts had been accorded his nearest male relative, Publius Aelius Hadrianus, the governor of Syria; and, according to Trajan's widow, Hadrian had actually been adopted by Trajan on his deathbed. Accordingly, both Senate and soldiers recognized him. Trajan's posthumous deification was never in doubt.

Hadrian (ruled 117–138), also a Spaniard, was an emperor of unusual versatility. Unlike Trajan, he was opposed to territorial expansion. Being himself in the East in 117, he renounced Trajan's conquests there immediately and contemplated evacuating Dacia as well. Furthermore, four of the consular generals particularly identified with Trajan's military ventures were arrested and executed "for conspiracy"; Hadrian claimed later that the Senate ordered their deaths against his wishes. His policies made his a peaceful reign. The only heavy fighting occurred in Judaea—or Syria Palaestina, as it was thenceforth called—

where Bar Kokhba led a furious, if futile, Jewish revolt (132–135) against Hadrian's conversion of Jerusalem into a Roman colony named Aelia Capitolina.

Hadrian's policies

Instead of expansion by war, Hadrian sought carefully delimited but well-defended frontiers, with client states beyond them where possible. The frontiers themselves, when not natural barriers, were strongly fortified: in Britain, Hadrian's Wall, a complex of ditches, mounds, forts, and stone wall, stretched across the island from the Tyne to the Solway; Germany and Raetia had a *limes* (fortified boundary) running between Mainz on the Rhine and Regensburg on the Danube. Within the frontiers the army was kept at full strength, mostly by local recruiting of legionaries and apparently of auxiliaries, too (so that Vespasian's system of having the latter serve far from their homelands gradually ceased). Moreover, the tendency for auxiliaries to be assimilated to legionaries continued; even the officers became less distinguishable, because equites now sometimes replaced senators in high posts in the legions. To keep his essentially sedentary army in constant readiness and at peak efficiency (no easy task), Hadrian carried out frequent personal inspections, spending about half his reign in the provinces (121–125; 128–134). The transformation of field troops into static garrisons frequently meant entrusting mobile operations to the so-called *numeri*—light-armed tribal units, often obtained outside the empire—and this contributed to the barbarization of the Roman imperial army.

Hadrian was also responsible for significant developments on the civilian side. Under him, equites were no longer required to do military service as an essential step in their career, and many of them were employed in the imperial civil service, more even than under Domitian. By now the formative days of the civil service were over; its bureaucratic phase was beginning, and it offered those equites who had no military aspirations an attractive, purely civilian career. Formal titles now marked the different equestrian grades of dignity: a procurator was *vir egregius;* an ordinary prefect, *vir perfectissimus;* a praetorian prefect, *vir eminentissimus,* the latter title being obviously parallel to the designation *vir clarissimus* for a senator. Thenceforth, equites replaced freedmen in the imperial household and bureaus, and they even appeared in Hadrian's imperial council.

Hadrian also improved legal administration. He had his expert jurists codify the *edictum perpetuum* (the set of rules gradually elaborated by the praetors for the interpretation of the law). He also appointed four former consuls to serve as circuit judges in Italy. This brought Italy close to a province; not that Hadrian was anxious to reduce the status of Italy—his policy was to make all parts of the empire important. For one part he was exceptionally solicitous: he spent much time in Greece and lavishly embellished Athens.

Hadrian maintained good relations with but was never fully trusted by the Senate. His foreign policy seemed to be unheroic, his cosmopolitanism to be un-Roman, and his reforms to encroach on activities traditionally reserved to senators. Moreover, in his last two years he was sometimes capricious and tyrannous. Like Augustus, he had no son of his own and conducted a frustrating search for a successor. After executing his only male blood relative, his grandnephew, in 136, he adopted Lucius Ceionius Commodus, renaming him Lucius Aelius Caesar. The latter, however, died shortly afterward, whereupon Hadrian in 138 chose a wealthy but sonless senator, the 51-year-old Titus Aurelius Antoninus; but, evidently intent on founding a dynasty, he made Antoninus in his turn adopt two youths, 16 and seven years old, respectively—they are known to history as Marcus Aurelius (the nephew of Antoninus' wife) and Lucius Verus (the son of Aelius Caesar). When Hadrian died soon thereafter, Antoninus succeeded and induced a reluctant Senate to deify the deceased emperor. According to some, it was this act of filial piety that won for Antoninus his cognomen, Pius.

Antoninus' reign of quiet prosperity

Antoninus Pius (ruled 138–161) epitomizes the Roman Empire at its cosmopolitan best. He himself was of Gallic origin; his wife was of Spanish origin. For most men his was a reign of quiet prosperity, and the empire under him deserves the praises lavished upon it by the contemporary writer Aelius Aristides. Unlike Hadrian, Antoninus travelled little; he remained in Italy, where in 148 he celebrated the 900th anniversary of Rome. *Princeps* and Senate were on excellent terms, and coins with the words *tranquillitas* and *concordia* on them in Antoninus' case mean what they say. Other of his coins not unreasonably proclaim *felicitas temporum* ("the happiness of the times"). Yet raids and rebellions in many of the borderlands (in Britain, Dacia, Mauretania, Egypt, Palaestina, and elsewhere) were danger symptoms, even though to the empire at large they seemed only faraway bad dreams, to use the expression of Aelius Aristides. Antoninus prudently pushed the Hadrianic frontiers forward in Dacia, the Rhineland, and Britain (where the Antonine Wall from the Forth to the Clyde became the new boundary) and carefully groomed his heir apparent for his imperial responsibilities.

Marcus Aurelius (ruled 161–180) succeeded the deified Antoninus and more than honoured Hadrian's intentions by immediately co-opting Lucius Verus as his full co-emperor. Because Verus' competence was unproved, this excess of zeal was imprudent. Fortunately, Verus left decision making to Marcus. Marcus' action was also dangerous for another reason; it represented a long step away from imperial unity and portended the ultimate division of the empire into Greek- and Latin-speaking halves. Nor was this the only foreboding development in Marcus' reign—formidable barbarian assaults were launched against the frontiers, anticipating those that were later to bring about the disintegration of the empire. Marcus himself was a Stoic philosopher; and his humanistic if somewhat pessimistic *Meditations* reveal how conscientiously he took his duties. Duty called him to war; he responded to the call and spent far more of his reign in the field than had any previous emperor.

At Marcus' very accession the Parthians turned aggressive, and he sent Verus to defend Roman interests (162). Verus greedily took credit for any victories but left serious fighting to Avidius Cassius and the army of Syria. Cassius succeeded in overrunning Mesopotamia and even took Ctesiphon, the Parthian capital; he was therefore able to conclude a peace that safeguarded Rome's eastern provinces and client kingdoms (166). In the process, however, his troops became infected with plague, and they carried it back with them to the west with calamitous results. The Danube frontier, already weakened by the dispatch of large detachments to the East, collapsed under barbarian assault. Pressed on from behind by Goths, Vandals, Lombards, and others, the Germanic Marcomanni and Quadi and the Sarmatian Iazyges poured over the river; the Germans actually crossed Raetia, Noricum, and Pannonia to raid northern Italy and besiege Aquileia. Marcus and Verus relieved the city shortly before Verus' death (169). Then, making Pannonia his pivot of manoeuvre, Marcus pushed the invaders back; by 175 they were again beyond the Danube. At that moment, however, a false report of Marcus' death prompted Avidius Cassius, by now in charge of all eastern provinces, to proclaim himself emperor. Cassius was himself a native Syrian, and his action may represent incipient separatism by the Greek-speaking East. In any case, it undid Marcus' achievements along the Danube because it took him to the East and reopened the door to barbarian attacks. Fortunately, Cassius was soon murdered, and Marcus could return to central Europe (177). But he had barely restored the frontier again when he died at Vindobona (Vienna) in 180, bequeathing the empire to his son, the 19-year-old Commodus, who had actually been named co-emperor three years earlier.

Barbarian assaults

Commodus (ruled 180–192), like Gaius and Nero, the youthful emperors before him, proved incompetent, conceited, and capricious. Fortunately, the frontiers remained intact, thanks to able provincial governors and to barbarian allies, who had been settled along the Danube with land grants and who gave military service in return. But Commodus abandoned Marcus' scheme for new trans-Danubian provinces, preferring to devote himself to sensual pleasures and especially to the excitements of the arena in Rome, where he posed as Hercules Romanus and forced the Senate to recognize his godhead officially. He

left serious business to his favourites; and their ambitions and intrigues led to plots, treason trials, confiscations, and insensate murders. Commodus' assassination on the last day of 192 terminated a disastrous reign; and the Antonines, like the Julio-Claudians, came to an ignominious end. And there was a similar sequel. Commodus' *damnatio memoriae,* like Nero's, was followed by a year of four emperors.

The empire in the 2nd century. The century and three-quarters after Augustus' death brought no fundamental changes to the principate, although so long a lapse of time naturally introduced modifications and shifts of emphasis. By Flavian and Antonine times the principate was accepted universally. For the provinces, a return to the republic was utterly unthinkable; for Rome and Italy, the year 69 served as a grim warning of the chaos to be expected if, in the absence of a *princeps,* the ambitions of a few powerful individuals obtained unfettered scope. A *princeps* was clearly a necessity, and men were even prepared to tolerate a bad one, although naturally they always hoped for a good one.

Nor did the *princeps* have to be chosen any longer from the Julio-Claudians. The great achievement of the Flavians was to reconcile the soldiers and the upper classes everywhere to the idea that others were eligible. The Flavians' frequent tenure of consulship and censorship invested their middle-class Italian family with the outward trappings of prestige and the aristocratic appearance of an authentic imperial household. The deification of the first two Flavians contributed to the same end, and so did the disappearance of old republican families that might have outranked the reigning house (by 69 most descendants of the republican nobility had either died of natural causes or been exterminated by imperial persecution). After the Flavians, the newness of a man's senatorial dignity and the obscurity of his ultimate origin, whether it was Italian or otherwise, no longer forbade his possible elevation. Indeed, Domitian's successors and even Domitian himself in his last years did not need to enhance their own importance by repeated consulships. The Antonine emperors, like the Julio-Claudians, held the office infrequently. They did, however, continue the Flavian practice of emphasizing the loftiness of their families by deifying deceased relatives (Trajan deified his sister, his niece, and his father; Antoninus, his wife; etc.).

Eligibility for the principate

Trend to absolute monarchy. Glorification of the reigning house, together with a document such as Vespasian's Lex de Imperio, helped to advertise the emperor's position; and under the Flavians and Antonines the principate became much more like an avowed monarchy. Proconsular *imperium* began to be reflected in the imperial titulary, and official documents started calling the emperor *dominus noster* ("our master").

The development of imperial law-making clearly illustrates the change. From the beginnings of the principate, the emperor had had the power to legislate, although no law is known that formally recognized his right to do so; by Antonine times, legal textbooks stated unequivocally that whatever the emperor ordered was legally binding. The early emperors usually made the Senate their mouthpiece and issued their laws in the form of senatorial decrees; by the 2nd century the emperor was openly replacing whatever other sources of written law had hitherto been permitted to function. After 100 the Assembly never met formally to pass a law, and the Senate often no longer bothered to couch its decrees in legal language, being content to repeat verbatim the speech with which the ruler had advocated the measure in question. After Hadrian, magistrates ceased modifying existing law by their legal interpretations because the praetors' *edictum perpetuum* had become a permanent code, which the emperor alone could alter. By 200, learned jurists had lost the right they had enjoyed since the time of Augustus of giving authoritative rulings on disputed points (*responsa prudentium*). Meanwhile, the emperor more and more was legislating directly by means of edicts, judgments, mandates, and rescripts—collectively known as *constitutiones principum.* He usually issued such *constitutiones* only after consulting the "friends" (*amici Caesaris*) who composed his imperial council. But a *constitutio* was nevertheless a fiat. The road to the later dominate (after 284) lay open.

Political life. Nevertheless, the autocratic aspect of the Flavian and Antonine regimes should not be overstressed. Augustus himself had been well aware that it was impossible to disguise permanently the supremacy that accumulation of powers gained piecemeal conferred; his deportment in his last years differed little from that of Vespasian, Titus, and the so-called five good emperors who followed them. Nor had other Julio-Claudians hesitated to parade their predominance—Claudius, by centralizing the imperial powers, reduced their apparent diversity to one all-embracing *imperium;* Gaius and Nero revealed the autocracy implicit in the principate with frank brutality.

What impresses perhaps as much as the undoubtedly autocratic behaviour of the Flavians and Antonines is the markedly civilian character of their reigns. They held supreme power, and some of them were very distinguished soldiers; yet they were not military despots. For this the old republican tradition—that a state official might serve it in both a civilian and a military capacity—was largely responsible. Matters could change after Hadrian separated the two roles. Actually, the 3rd century soon showed what it meant to have a *princeps* whose whole experience had been confined to camps and barracks.

Civilian character of the Flavian and Antonine reigns

As imperial powers became more concentrated, republican institutions decayed; the importance of imperial officials grew, while the authority of urban magistrates declined. Quaestorship, praetorship, and consulship (the last named now reduced to a two-month sinecure) became mere stepping stones to the great imperial posts that counted most in the life of the empire. Governors of imperial provinces and commanders of legions were Roman senators; but they were equally imperial appointees. Clearly, the emperor was the master of the Senate; and it was disingenuous for him to get impatient, as some emperors did, with the Senate's lack of initiative and reluctance to take firm decisions of its own. The emperor might not even consult the Senate much, preferring to rely on his imperial council, which by the 2nd century was no longer composed exclusively of senators.

The Senate, however, while exercising little power, was treated courteously by most Flavians and Antonines. They recognized its importance as a lawcourt, as the body that formally appointed a new emperor, and as a sounding board of informed opinion. Senators came increasingly from the provinces, and, although this meant preeminently the western provinces (the Greek-speaking East being underrepresented), the Senate did reflect to some extent the views of the empire at large.

The equites, meanwhile, steadily acquired greater importance as imperial officials. In newly created posts they were invariably the incumbents, and in posts of long standing they replaced freedmen and *publicani.* During the 2nd century equestrian procurators became the solid basis of the civil service—four grades, distinguished by salary, were established. The great bureaucracy of later times was fairly launched. This development inevitably bred bitter rivalry with the Senate, mitigated somewhat by the similar social backgrounds of the two orders.

Rome and Italy. By the 2nd century the city of Rome had attracted freeborn migrants from all over the empire; it housed, additionally, large numbers of manumitted slaves. These newcomers were all assimilated, making Rome a veritable melting pot, and they hoped—especially the many who came from the East—to dilute the city's Italian flavour. The vast majority of them were poor, a handful of opulent imperial freedmen being entirely exceptional. But many were energetic, enterprising, and lucky and made their way in the world. Freedmen laboured under a social stigma, although some of them managed to become equites. Their sons, however, might overcome discrimination, and their grandsons were even eligible for membership in the Senate.

Rome as a melting pot

Inevitably, there was extensive trade and commerce (much of it in freedman hands) in so large a city, which was also the centre of imperial administration. There was little industry, however, and the urban poor had difficulty finding steady employment. Theirs was a precarious ex-

istence, dependent on the public grain dole and on the private charity of the wealthy. Large building programs gave Flavian and Antonine emperors the opportunity not only to repair the damage caused by fire and falling buildings (a frequent hazard among the densely packed and flimsily built accommodations for the urban plebs) but also to relieve widespread urban unemployment. This also made imperial Rome a city of grandeur. Augustus' building program had been vast but mostly concerned with repairing or rebuilding structures already existing, and his Julio-Claudian successors had built relatively little until the great fire made room for the megalomaniac marvels of Nero's last years. It was under the Flavians and Antonines that Rome obtained many of its most celebrated structures: the Colosseum, Palatine palaces, Trajan's Forum, the Pantheon, the Castel Sant' Angelo (Hadrian's mausoleum), the Temple of Antoninus and Faustina, Aurelius' Column, as well as the aqueducts whose arches marched across the Campagna to keep the city and its innumerable fountains supplied with water.

Italy was much less cosmopolitan and sophisticated and, according to literary tradition, much more sober and straitlaced than was Rome. It was the mistress of the empire, although the gap between it and the provinces was narrowing. Hadrian's policies especially helped to reduce its privileged position. His use of circuit judges was resented precisely because with them Italy resembled a province; actually, Italy badly needed them, and their abolition by Antoninus Pius was soon reversed by Marcus Aurelius. Also in Aurelius' reign a provincial fate overtook Italy in the form of barbarian invasion; and a few years later the country got its first legionary garrison under Septimius Severus.

The economic importance of Italy also declined. By Antonine times the eastern provinces had overtaken it, and it was hardly ahead even of those of the west. Gaul, for instance, could rival Italy industrially. In agricultural products, northern Africa and Spain are revealed by the potsherds of Monte Testaccio to have become serious competitors to Italian growers. The emperors, however, were not indifferent to Italy—Domitian, for example, protected Italian viticulture by restricting vine growing in the provinces; Trajan and his successors forced Roman senators to take an interest in the country, which was no longer the homeland of many of them, by investing a high proportion of their capital in Italian land (one-third under Trajan, one-quarter under Aurelius).

Developments in the provinces. In the empire at large, Flavians and Antonines, like the better Julio-Claudians, aimed at stability in order that its inhabitants might live in security and self-respect. In this they largely succeeded. Gibbon's famous description of the 2nd century as the period when men were happiest and most prosperous is not entirely false. Certainly, by then men had come to take for granted the unique greatness and invincible eternity of the empire; even the ominous events of Aurelius' reign failed to shatter their conviction that the empire was impregnable. The empire was a vast congeries of peoples and races with differing religions, customs, and languages, and the emperors were content to let them live their own lives. Imperial policy favoured a veneer of common culture transcending ethnic differences, but there was no deliberate denationalization. Ambitious men striving for a career naturally found it helpful, if not necessary, to become Roman in bearing and conduct and perhaps even in language as well (although speakers of Greek often rose to exalted positions). But local self-government was the general rule, and neither Latin nor Roman ways were imposed on the communities composing the empire. The official attitude to religion illustrates this—in line with the absolutist trend, emperor worship was becoming slowly but progressively more theocratic (Domitian relished the title of god, Commodus demanded it); yet this did not lead to the suppression of non-Roman or even outlandish cults, unless they were thought immoral (like Druidism, with its human sacrifice) or conducive to public disorder (like Christianity, with its uncompromising rejection of emperor worship and its liability to become a target for riots).

Attempts to achieve stability in the provinces

Where possible, the emperors kept direct administration from Rome to a minimum (except perhaps in Egypt), and the 2nd century was the most flourishing period of urban civilization that the empire ever knew. Administration everywhere was in the hands of the local well-to-do, who alone could afford the costs attaching to it. Local magistrates paid substantial fees on election; nor did their expenses end there, for they remained members of their local council after their term of office, and local councillors (*decuriones,* later *curiales*) were expected to contribute handsomely to local amenities. The cost involved inevitably reduced local elections to formalities, and the local councils became self-perpetuating bodies whose members were persons of some consequence. It was from these local worthies that the emperor often found his candidates for the Senate at Rome, an honour that was eagerly sought by individuals but that was a mixed blessing for their local communities, which stood thereby to lose prospective benefactors.

It is impossible not to be impressed by the spectacle of the Roman Empire in its 2nd-century heyday, with its panorama of splendid and autonomous communities. Nevertheless, before the century was over, there was growing difficulty in maintaining flourishing municipal life in a world where the ordinary man was encouraged to regard the emperor as a sort of terrestrial Providence and where the emperor himself with responsible earnestness accepted the role of universal dispenser of justice. The letters of the younger Pliny and of Marcus Cornelius Fronto reveal how seriously the 2nd-century emperors took their duty and strove for orderly government everywhere. But the emperors' very conscientiousness led inevitably to interference with local autonomy. Perhaps the civil service that Augustus founded would have burgeoned in any event and encroached on the self-governing communities that made up the provinces; but the well-meaning efforts of the Antonines hastened such a development. The ultimate effect was to dampen civic ardour and to foster listlessness. Faced with the prospect of increasing direction from above, municipal notables began avoiding local office. Inability to pay the cost involved may also in part explain the growing reluctance of men to undertake municipal responsibilities; although the provincial bourgeoisie remained generally prosperous, economic recession had set in before the 2nd century ended. For whatever reason, local officeholders became less easy to find; and, well before 200, men were being compelled to accept local office. This boded ill for the future.

The emperors' interference with local autonomy

Information about the lower orders throughout the empire is not very plentiful. So long as economic activity remained at a high level, there was apparently much social mobility. This was just as well, because there were great disparities of wealth and of class. Society was stratified, and the gap between poor and rich in provincial towns somewhat resembled that between plebs and aristocracy, senatorial or equestrian, in Rome. Indeed, it was precisely in Antoninus' reign that Roman law began to distinguish the lower classes (*humiliores*) from the upper (*honestiores*), and, although local customs and traditions were respected, the civil law of Rome was gradually replacing local codes. The 2nd century and the opening years of the 3rd comprised the classical period of Roman law, producing such illustrious jurists as Salvius Julianus, Gaius, Papinian, Ulpian, and Julius Paulus; and the temptation for the communities of the empire to adopt it must have been strong.

Clearly, Romanization was spreading and with it the Roman citizenship, especially from Hadrian's reign on. The number of provincials with citizenship increased so rapidly that 20 years after Commodus' death the emperor Caracalla could enfranchise all the communities of the empire and excite very little comment in doing so. The subjects of the empire found this levelling of status comfortable; they were also grateful for the peace that prevailed over much of the empire, but they were not particularly eager to defend it. Roman citizens in general avoided military service, and the empire obtained its soldiers by offering Roman citizenship as a reward for enlistment. The ever-wider extension of the citizenship ended its usefulness as

a bribe, and the recruiting areas steadily contracted to the least Romanized districts. Finally, recruits had to be obtained outside the empire altogether.

Evidently, the empire did not inspire a sense of duty or high resolve. There were both political and economic reasons. The average nonwealthy man hardly participated in political life, even in his own community, and his share of the prevailing prosperity was small. Slavery was ubiquitous; and, although humanitarian legislation gradually softened its harsher features, its unhappy victims depressed the living standards of the free workers so that poverty was a serious social problem. Drab materialism was the lot of many—in Asia Minor, for instance, despite unparalleled prosperity brought about by its having the whole Mediterranean open to it as at no other time, great unrest developed as a result of extremes of affluence and poverty. Many men sought solace from worldly cares in the mystery religions, and the prevailing polytheism enabled them to switch or multiply their gods without much problem of conscience (except for the increasingly numerous Christian converts).

Growing use of Latin in Iberia and Gaul

By the 2nd century considerable Latinization had occurred in the West. Modern Spanish, Portuguese, and French show this was particularly true of the Iberian Peninsula, which had been provincial soil ever since the Second Punic War, and of Gaul, where Latin enjoyed the advantage of some relationship to Celtic. In these regions even the lower orders adopted Latin, although the survival of Basque proves that the older native languages did not perish utterly. Moreover, enemy threats were far enough away and integration far enough advanced to render large garrisons unnecessary. Elsewhere, and especially in central Europe, large forces remained and were themselves potent instruments of Latinization, because in the army Latin was both the official language and also the lingua franca for heteroglot recruits. Furthermore, thanks to the *coloniae,* urbanization increased; and urbanization promoted Latinization. In the Danubian region, for instance, where the barbarian threat was acute, ceaseless military activity led to the establishment of numerous *coloniae,* to the spread of Latin (and its survival in modern Romanian), to the disappearance of ancient Illyrian and Thracian (unless modern Albanian preserves them), and to the birth of a strong pro-Roman patriotism, which increased men's eagerness to learn Latin. In areas that were some distance away from the frontiers, and especially in areas that were lacking the economic resources to generate cities, often only the bourgeoisie acquired Latin; among the lower orders the pre-Roman languages, even if seldom written, continued in use: Celtic in Britain and Switzerland and, to a lesser extent, in Gaul; Berber and some Punic in Africa and Mauretania. Moreover, from Hadrian on—when local recruiting and local service became the rule, and *coloniae* were no longer new foundations but simply old towns upgraded in status—the army lost much of its Latinizing role. Complete linguistic unity was therefore never achieved in the West.

This was true also of the East, where Greek had the status of an official language. Rome readily accepted and even encouraged Hellenization of the whole region; and, although the Romans were gratified if Orientals adopted Latin, they realized the futility of trying to impose it. In many districts, however, Greek was spoken only by the administrative and propertied classes, just like Latin in the West. Lower orders, especially in inland districts, continued to use their vernaculars: Aramaic in Syria, Coptic in Egypt, and a variety of tongues in Asia Minor.

The official emperor worship became universal, but Rome was no fountainhead of religious ideas; local cults persisted in both East and West, and the East even gave birth to new religions, such as Christianity. The native divinities were worshipped either under their own names or under those of Greco-Roman deities with whom they could be conveniently identified, for there was much syncretism. In religion as in language the army played an important role. The soldiers, with some help from eastern merchants, spread the worship of such deities as Mithra and Dolichenus, with whom they had become acquainted in the East, to the farthest corners of the empire; Mithra became known even in distant Britain, and Dolichenus was assimilated to Jupiter everywhere.

In the material aspects of life there was much Romanization. Towns in the Latin-speaking provinces looked very Roman in layout and in buildings: their streets were arranged on a checkerboard plan, and they usually boasted forums, triumphal arches, basilicas, baths, and Roman-type temples, theatres, and amphitheatres, graced with Roman sculpture and decoration. The eastern cities were much less receptive to Roman influences; in fact, the Latin West and the Greek East tended to go their separate ways. Differing outlooks and values, divergent interests and degrees of development, mutual suspicions and antipathies combined to keep them apart; and by Marcus Aurelius' day the blueprint for a divided empire had already been sketched in outline, no matter how faintly. Yet, even in the East, Roman influence was not negligible: some Greek cities became partly Roman in aspect, went in for gladiatorial games and similar Roman entertainments, and adopted the Roman civil law, the most enduring of the empire's legacies.

Differences between the Greek East and the Latin West

Economic life. Romanization was due chiefly to Rome's political authority, but the conditions fostered by the Pax Romana also helped. Harbour dues and customs duties prevented the Roman Empire from being one large free-trade area. Nor was there complete laissez faire in other respects, despite a general lack of policies for improving the increasingly complex economy by promoting trade and regulating commerce: there was much routine administration directly concerned with production and distribution of goods. The supervision of the enormous and ubiquitous imperial properties, the provision of Rome's food supply, and procurement for the armed services and other branches of government constituted a significant proportion of the empire's economic activity. Nevertheless, in the prevailing peace, the private sector offered great scope for initiative, enterprise, the free play of competition, and general economic development. Industrial and commercial methods, banking, and credit facilities were all improved, and goods moved easily along the highways crisscrossing the empire from frontier to frontier. There was some trade with countries beyond the frontiers, such as Scandinavia and even China (indirectly), although naturally there was far more within the empire itself. In the West, large Roman (or at least Romanized) towns came into being to serve the needs of trade. In the East, Asia Minor and Syria interchanged goods with the whole Mediterranean world, and the Greek cities flourished as never before, Alexandria and Antioch being the greatest of them but with those of Asia Minor not far behind.

Waning of prosperity during the 2nd century

But in economic as in other matters there was premonition of trouble to come. Before the 2nd century ended, prosperity had begun to wane. A clear symptom of this was the reduction in the weight and fineness of the standard silver coin, the denarius. The economy faltered for various reasons: the consistent drain of metal to the East to pay for spices and Oriental luxuries, the maldistribution of wealth, the lack of industrial inventiveness in a slave-based society, the depopulation caused by war (by Aurelius' reign no longer confined to frontier areas). The economic chaos of the 3rd century, however, was still some distance away.

The army. The army that enforced the Pax Romana had expanded little beyond the size envisaged for it by Augustus, despite the enlargement of the empire by Claudius, the Flavians, and Trajan. It reached 31 legions momentarily under Trajan, but it usually totalled 28 under the Flavians and Antonines until the onset of the frontier crisis in Aurelius' reign brought it to 30. This, seemingly, was the limit for which either recruits or money could be found; and it probably explains why Hadrian, and later Commodus, halted further expansion.

The army was used not to prop up a militarist government but to defend the frontiers. Shifts in enemy pressures, however, caused the legions to be distributed differently than in Julio-Claudian times. Under Antoninus Pius, the Danubian provinces (Pannonia, Moesia, Dacia) had ten, and the East (Asia Minor, Syria, Palestine, Egypt) had nine, and both regions also had supporting naval flotillas;

of the remaining nine legions, Britain contained three and the Rhineland four. Clearly, the army was thinly stretched elsewhere, with auxiliaries bearing the burden of defense. In much of the empire the sight of a soldier, especially a legionary, was a comparatively rare event, the more so because normally troops not only remained in frontier areas while serving but were also largely recruited there.

The provinces underwent changes in size as well as in garrison. For prudential reasons the emperors refused to entrust many legions to a single commander; they split a province in two as soon as it needed a large military establishment. Thus, the number of provinces was constantly growing; by Hadrian's day subdivision began to anticipate the fragmentation later carried out by Diocletian.

Imperial literature

Cultural life. The literature of the empire is both abundant and competent, for which the emperors' encouragement and financing of libraries and higher education were perhaps in part responsible. The writers, however, with the possible exception of Christian apologists, were seldom excitingly original and creative. As Tacitus said, the great masters of literature have ceased to be. Perhaps Augustus' emphasis on tradition affected more than political ideals and practice. At any rate, men of letters, too, looked often backward. At the same time they clearly reveal the success of the empire in spreading Greco-Roman culture, for the majority of them were natives of neither Italy nor Greece. Of the writers in Latin, the two Senecas, Lucan, Martial, Columella, Hyginus, and Pomponius Mela came from Spain; Fronto, Apuleius, and probably Florus and Aulus Gellius, from Africa. Tacitus was perhaps from Gallia Narbonensis. The Latin writers in general sought their models less in Greece than in Augustus' Golden Age, when Latin literature had reached maturity. Thus, the poets admired Virgil and imitated Ovid; lacking genuine inspiration, they substituted for it an erudite cleverness, the fruit of an education that stressed oratory of a striking but sterile kind. Authentic eloquence in Latin came to an end when, as Tacitus put it, the principate "pacified" oratory. Under the Flavians and Antonines, an artificial rhetoric, constantly straining after meretricious effects, replaced it. The epigrammatic aphorism (*sententia*) was especially cultivated; the epics of Lucan, Valerius Flaccus, Silius Italicus, and Statius are full of it, and it found a natural outlet in satirical writing, of which the Latin instinct for the mordant always ensured an abundance. In fact, Latin satire excelled: witness Martial's epigrams, Petronius' and Juvenal's pictures of the period, and Persius' more academic talent. For that matter, Tacitus' irony and pessimism were not far removed from satire.

In the East the official status of Greek and the favour it enjoyed from such emperors as Hadrian gave new life to Greek literature. It had something in common with its Latin counterpart in that it looked to the past but was chiefly written by authors who were not native to the birthplace of the language. The so-called Second Sophistic reverted to the atticism of an earlier day but often in a Roman spirit; its products from the Asian pens of Dio Chrysostom and Aelius Aristides are sometimes limpid and talented tours de force but rarely great literature. In Greek, too, the best work was in satire, the comic prose dialogues of the Syrian Lucian being the most noteworthy and original literary creations of the period. Among minor writers the charm of Arrian and Pausanias, Asians both, and above all of Plutarch abides (although Plutarch's talents were mediocre, and his moralizing was shallow, his biographies, like those of his Latin contemporary Suetonius, are full of information and interest).

Imperial encouragement of Greek culture and a conviction, no longer justified, of its artistic and intellectual superiority caused the East to resist Latinization. This attitude was bound to lead to a divided empire, and thoughtful observers must have noted it with misgivings. The split, however, was still far in the future. Meanwhile, there was a more immediate cause for disquiet. The plethora of summaries and anthologies that appeared implies a public progressively indifferent to reading. In other words, the outlook for letters was poor, and this had an unfortunate effect on the scientific literature of the age, which was in itself of first-class quality. Dioscorides on botany, Galen on medicine, and Ptolemy on mathematics, astronomy, and geography represent expert scholars, expounding carefully, systematically, and lucidly the existing knowledge in their respective fields. But their very excellence proved fatal because, as the reading public dwindled, theirs remained standard works for far too long; their inevitable errors became enshrined, and their works acted as brakes on further progress.

Stoicism was the most flourishing philosophy of the age. In the East a sterile scholasticism diligently studied Plato and Aristotle, but Epictetus, the Stoic from Asia Minor, was the preeminent philosopher. In the West, Stoicism permeates Seneca's work and much of Pliny's *Natural History.* Evidently, its advocacy of common morality appealed to the traditional Roman sense of decorum and duty, and its doctrine of a world directed by an all-embracing providence struck a responsive chord in the 2nd-century emperors, though they deeply disapproved of its extremist offshoots, the Cynics: Marcus Aurelius was himself a Stoic.

Imperial art

Imperial art, dealing above all with man and his achievements, excelled in portraits and commemoration of events; Roman sculpture and presumably Roman painting, also, owed much to Greek styles and techniques. It emerged, however, as its own distinctive type. The Augustan age had pointed the way that Roman art would go: Italian taste would be imposed on Hellenic models to produce something different and original. The reliefs of the Augustan Ara Pacis belong to Rome and Italy, no matter who actually carved them. By Flavian times this Roman instinct had asserted itself and with it the old Roman tendency toward lively and accurate pictorial representation. This can be seen from the reliefs illustrating the triumph over Judaea in the passageway of the Arch of Titus. The narrative description dear to Roman art found its best expression in the great spiral frieze on Trajan's Column, where the Emperor can be seen among his soldiers at various times in the Dacian campaigns; the story of the war plays a most important part, although, like most imperial monuments, the column is meant to exalt the leader. Under Hadrian a reaction made sculpture less markedly Italian, as if to conform with the slow decline of Italy toward quasi-provincial status. Also under Hadrian, the emperor was more singled out, being made bigger and more frontal than the other figures, as if to illustrate the growing monarchical tone of the principate. This tendency continued under the Antonines, when there was a magnificent flowering of sculpture on panels, columns, and sarcophagi; but its exuberance and splendour foreshadow the end of classical art.

The artistic currents that flowed in Rome were felt throughout the empire, the less developed areas being influenced most. In the West, provincial sculpture closely resembled Roman, although it sometimes showed variations, in Gaul especially, due to local influences (the native element, however, is not always easy to identify). The Roman quality of portraits painted on Egyptian mummy cases shows that the Greek-speaking regions were also affected, although generally they maintained their own traditions. But by now the Greek East had become rather barren; much of its production was imitative rather than vitally creative. Greece proper contributed little, the centre of Hellenism having shifted to Asia Minor, to places like Aphrodisias, where there was a flourishing school of sculpture.

In at least one respect the East was heavily influenced by Rome. The use of concrete and cross vault enabled Roman architects and engineers to span wide areas; their technological achievements included the covered vastness of the huge thermal establishments, the massive solidity of the amphitheatres, and the audacity of the soaring bridges and aqueducts. The East was greatly impressed. Admittedly, the agoras and gymnasiums in Greek towns are hardly Roman in aspect, but, for most structures of a practical utilitarian kind, the Greek debt to Rome was heavy. Sometimes Roman influence can be seen not only in the fundamental engineering of such buildings as market gateways, theatres, and amphitheatres, but even in such decorative details as composite capitals as well.

Roman features abound in exotic Petra, Palmyra, Gerasa, and Baalbek, and even in Athens itself. (E.T.S.)

THE LATER ROMAN EMPIRE

The dynasty of the Severi (AD 193–235). *Septimius Severus.* After the assassination of Commodus on December 31, AD 192, Helvius Pertinax, the prefect of the city, became emperor. In spite of his modest birth, he was well respected by the Senate but was without his own army. He was killed by the praetorians at the end of March 193, after a three-month reign. The praetorians, after much corrupt bargaining, designated as emperor an old general, Didius Julianus, who had promised them the largest *donativum* (a donation given to each soldier on the emperor's accession). The action of the praetorians roused the ire of the provincial armies. The army of the Danube, which was the most powerful as well as the closest to Rome, appointed Septimius Severus in May 193. Severus soon had to face two competitors, supported, like himself, by their own troops: Pescennius Niger, the legate of Syria; and Clodius Albinus, legate of Britain. After having temporarily neutralized Albinus by accepting him as Caesar, Septimius marched against Niger, whose troops, having come from Egypt and Syria, were already occupying Byzantium. The Danubian legions were victorious, and Niger was killed at the end of 194; Antioch and Byzantium were pillaged after a long siege. Septimius even invaded Mesopotamia, for the Parthians had supported Niger. But this campaign was quickly interrupted: in the West, Albinus, disappointed at not being associated with the empire, proclaimed himself Augustus in 196 and invaded Gaul. He was supported by the troops, by the population, and even by the senators in Rome. In February 197, he was defeated and killed in a difficult battle near his capital of Lyons which, in turn, was almost devastated. Septimius Severus remained the sole master of the empire, but the pillagings, executions, and confiscations left a painful memory. A few months later, in the summer of 197, he launched a second Mesopotamian campaign, this time against the Parthian king Vologases IV, who had attacked Nisibis, which the Romans had conquered two years previously. Septimius Severus was again victorious. Having arrived at the Parthian capitals (Seleucia and Ctesiphon), he was defeated near Hatra but in 198 obtained an advantageous peace: Rome retained a part of Mesopotamia, together with Nisibis, the new province being governed by an eques. After having inspected the East, the Emperor returned to Rome in 202. He spent most of his time there until 208, when the incursions of Caledonian rebels called him to Britain, where he carried out a three-year campaign along Hadrian's Wall. He died at Eboracum (York) in February 211.

Severus' campaigns

Septimius Severus belonged to a Romanized Tripolitan family that had only recently attained honours. He was born in Leptis Magna in North Africa and favoured his native land throughout his reign. Married to Julia Domna of Emesa, a Syrian woman from an important priestly family, he was surrounded by Easterners. He had pursued a senatorial career and had proved himself a competent general, but he was above all a good administrator and a jurist. Disliking Romans, Italians, and senators, he deliberately relied on the faithful Danubian army that had brought him to power and always showed great concern for the provincials and the lower classes. Although he had sought to appropriate the popularity of the Antonines to his own advantage by proclaiming himself the son of Marcus Aurelius and by naming his own son Marcus Aurelius Antoninus, he in fact carried out a totally different policy—a brutal, yet realistic policy that opened careers to new social classes. Indifferent to the prestige of the Senate, where he had a great many enemies, he favoured the equites. The army thus became the seedbed of the equestrian order and was the object of all of his attentions. The ready forces were increased by the creation of three new legions commanded by equites; and one of these, the Second Parthica, was installed near Rome. Unlike Vespasian, who also owed his power to the army but who knew how to keep it in its proper place, Septimius Severus, aware of the urgency of external problems, established a sort of military monarchy. The praetorian cohorts doubled their ranks, and the dismissal of the old staff of Italian origin transformed the Praetorian Guard into an imperial guard in which the elite of the Danube army were the most important element. The auxiliary troops were increased by the creation of 1,000-man units (infantry cohorts) and cavalry troops, sometimes outfitted with mail armour in the Parthian manner. The careers of noncommissioned officers emerging from the ranks now opened onto new horizons: centurions and noncommissioned grades could attain the tribunate and enter into the equestrian order. Thus, a simple Illyrian peasant might attain high posts: this was undoubtedly the most essential aspect of the "Severan revolution." This "democratization" was not necessarily a barbarization, for the provincial legions had long been Romanized. Their salaries were increased, and *donativa* were distributed more frequently; henceforth, soldiers were fed at the expense of the provincials. Veterans received lands, mostly in Syria and Africa. The right of legitimate marriage, previously refused by Augustus, was granted to almost all of the soldiers, and the right to form *collegia* (private associations) was given to noncommissioned officers. The army was tending to become a privileged caste, issuing from the peasantry. In the 1920s, the Russian scholar Michael Rostovtzeff suggested that this policy marked the beginning of a socially dangerous collusion between the peasants and the army—which was recruited from among them—and indicated a class struggle against the bourgeoisie of the city. But many scholars now believe that Rostovtzeff, perhaps influenced by parallels with the Red Army during the Russian Revolution, overstated his case.

The growth of the equestrian order

The administrative accomplishments of Septimius Severus were of great importance: he clearly outlined the powers of the city prefect; he entrusted the praetorian prefecture to first class jurists, such as Papinian; and he increased the number of procurators, who were recruited for financial posts from among Africans and Easterners and for government posts (*praesides*) from among Danubian officers. Italy lost its privileges and found itself subjected, like all the other provinces, to the new *annona,* a tax paid in kind, which assured the maintenance of the army and of the officials. The consequent increase in expenditures—for administration, for the salaries and the *donativa* of the soldiers, for the maintenance of the Roman plebs, and for construction—obliged the Emperor to devalue the denarius in 194. But the confiscations increased his personal fortune, the *res privata,* which had been previously created by Antoninus.

Severus' social policy favoured both the provincial recruitment of senators (Easterners, Africans, and even Egyptians), causing a sharp decrease in the percentage of Italian senators, and the elevation of the equestrian order, which began to fill the prince's council with its jurists. The cities, which had been favoured by the Antonines, were more and more considered as administrative wheels in the service of the state: the richest *decuriones* (municipal councillors) were financially responsible for levying the taxes, and it was for this purpose that the towns of Egypt finally received a *boulē* (municipal senate).

The burden of taxes and forced government service was made weightier by numerous transport duties for the army and for the *annona* service and was regulated by the jurists through financial, personal, or mixed charges. The state was watchful to keep the *decuriones* in the service of their cities and to provide a control on their administration through the appointment of *curatores rei publicae,* or officials of the central government. The lower classes were, in principle, protected against the abuses of the rich, but in fact they were placed at the service of the state through the restrictions imposed on shipping and commercial corporations. The state became more and more a policeman, and the excesses of power of numerous grain merchants (*frumentarii*) weighed heavily on the little man.

Imperial power, without repudiating the ideological themes of the principate, rested in fact on the army and sought its legitimacy in heredity: the two sons of Septimius Severus, Caracalla and Geta, were first proclaimed Caesars (heirs apparent), the former in 196, the latter in 198; later, they were directly associated in imperial power

through bestowal of the title of Augustus, in 198 and 209, respectively. Thus, during the last three years of Septimius Severus' reign, the empire had three Augusti at its head.

Caracalla. Caracalla, the eldest son of Septimius Severus, reigned from 211 to 217, after having assassinated his younger brother, Geta. He was a caricature of his father: violent, megalomaniacal, full of complexes, and, in addition, cruel and debauched. He retained, however, the entourage of the equites and jurists who had governed with his father and enforced to an even greater degree his father's militaristic and egalitarian policy. He increased the wages of the army even further and, at the same time, began a costly building program that quickly depleted the fortune left him by his father. He forced the senators to pay heavy contributions, doubled the inheritance and emancipation taxes, and often required the *aurum coronarium* (a contribution in gold), thereby ruining the urban middle classes. To stop the inflation that was depreciating the silver denarius, he created a new coin, the *antoninianus,* whose real value was that of the Marcus Aurelius coin but whose official rate was fixed at twice the value, thus insuring that monetary instability would continue for some time to come. In 212, the famous Constitutio Antoniniana de Civitate gave the right of citizenship to nearly all inhabitants of the empire. This law, known only through a mutilated papyrus, has raised many controversies; but even taking into account the exclusion of numerous barbarians and of a part of the indigenous population of Egypt, the effect of this measure was considerable. It sanctioned the political evolution of several centuries and represented true social advancement for many provincials. It also reinforced imperial unity and complied with the wishes of the great jurists of the time. Each inhabitant of the empire would henceforth have a double citizenship—that of Rome and that of his city of origin. The very diffusion, however, of citizenship status diminished its value, and, in effect, all citizens were reduced to subjects of an absolute prince. On the other hand, the recruitment of legionaries and of municipal magistrates became more difficult, for up to that time acquisition of Roman citizenship had been the principal inducement for undertaking such offices.

Extension of Roman citizenship

Although little endowed with military qualities, Caracalla adopted as his patron Alexander the Great, whom he admired greatly, and embarked on an active external policy. He fought successfully against the Teutonic tribes of the Upper Danube, among whom the Alamanni, as well as the Capri of the Middle Danube, appeared for the first time; he often prudently mixed military operations with negotiation and gave important subsidies and money (in sound currency) to the barbarians, thus arousing much discontent. His ambition was to triumph in the East like his hero of old and, more recently, Trajan and his own father. He invaded Armenia and Adiabene and annexed Osroëne in northwest Mesopotamia, joining it to the part of Mesopotamia taken by Septimius Severus. In April 217, while pursuing his march on the Tigris, he was assassinated on the order of one of his praetorian prefects, Marcus Opellius Macrinus.

Macrinus. Macrinus was accepted as emperor by the soldiers, who were unaware of the role he had played in the death of his predecessor. For the first time an eques had acceded to the empire after having been no more than a manager of financial affairs. The senators reluctantly accepted this member of the equestrian order, who, nevertheless, proved to be moderate and conciliatory; but the armies despised him as a mere civilian, and the ancient authors were hostile to him. His reign was brief, and little is known of him. He concluded an inglorious peace with the Parthians, which assured Mesopotamia to Rome through the payment of large sums of money. And to make himself popular, he cancelled Caracalla's tax increases and reduced military expenditures. A plot against him was soon organized: two young grandnephews of Septimius Severus were persuaded by their mothers and especially by their grandmother, Julia Maesa, the sister of Julia Domna (who had recently died), to reach for imperial power. The eldest, Bassianus, was presented to the troops of Syria, who had been bought with gold, and was proclaimed in April 218. Shortly afterward, Macrinus was defeated and killed, as was his son (whom he had associated with him on the throne).

Elagabalus and Severus Alexander. The new emperor was presented as the son of Caracalla, whose name he took (Marcus Aurelius Antoninus). He is better known, however, under the name Elagabalus, the god whose high priest he was and whom he quickly and imprudently attempted to impose on the Romans, in spite of his grandmother's counsel of moderation. Fourteen years old, he caused himself to be detested by his heavy expenditures, his orgies, and the dissolute behaviour of his circle. His fanatic devotion to the Baal of Emesa (Elagabal) was sincere, but he shocked the Romans by his extravagances. The praetorians killed him in 222 and proclaimed as emperor his first cousin, Alexianus, who took the name of Severus Alexander.

Although well educated and full of good intentions, Severus Alexander showed some weakness of character by submitting to the counsel of his mother, Mamaea, and of his grandmother, Maesa. The *Scriptores historiae Augustae,* a collection of biographies of the emperors, attributes to him a complete program of reforms favourable to the Senate, but these reforms are not mentioned elsewhere. As in the time of Septimius Severus, his counsellors were equites. Ulpian, the praetorian prefect, was the greatest jurist of this period, and the basic policies of the founder of the dynasty were carried on but with less energy. This weakening of energy had disastrous results: in Persia, the Arsacids were replaced in 224 by the more ambitious Sāsānid dynasty, who hoped to recover the former possessions of the Achaemenids in the East. Their initial attacks were stopped in 232 by a campaign that was, however, poorly conducted by the Emperor and that alienated the army as a result of its ineptitude. In Rome, there were frequent disorders and, as early as 223, Ulpian had been killed by the praetorians. While gathered on the Rhine to fight the Teutons, the soldiers once again revolted and killed Severus Alexander and his mother. A coarse and uneducated but energetic soldier, Maximinus the Thracian, succeeded him without difficulty in March 235. The Severan dynasty had come to an end.

Rise of the Sāsānids in Persia

Religious and cultural life in the 3rd century. *Official religions and Eastern cults.* Since the reforms of Augustus and his efforts to restore the traditional religion, few religious changes had taken place. The official ceremonies in honour of the Capitoline divinities Jupiter, Juno, and Minerva, the protectors of the state, were still celebrated by the priests of the sacerdotal *collegia* of Rome, of the provinces, and of the municipalities, but their religious vigour was waning. In the countryside, local agrarian and fertility cults survived, and in the course of the 2nd century AD there was a renaissance of the ancient indigenous cults in the Celtic countries of Gaul and Britain, in Africa, and in Central Europe. Rather than the great gods, who were too "political," the lower classes preferred the more helpful divinities, such as Hercules and Silvanus. Simplistic superstitions still flourished, and even the educated classes did not hesitate to indulge in the magical practices that the law sometimes condemned. The 2nd-century philosopher and rhetorician Apuleius, for example, was brought to trial for having wooed a rich widow by magic arts.

The cult of the emperor had deeply penetrated the traditional religion; and there was a tendency, especially in the East, to make divinities of the emperors while they were still alive. The Severan period was characterized by the triumph of the Oriental religions and by the first efforts of syncretism. For some time now, the Eastern religions had spoken successfully to the uneasy soul of the masses. To be sure, the Greek gods Demeter and Dionysus still had their mysteries and retained their followers. But the cults of the East were aided by their very strangeness, by the fervour of their ceremonies, and by the hope of a future life they reserved to their initiates, who were fraternally united around a well-organized clergy. The oldest established religions were those of Cybele (the Phrygian Great Mother) and of her ancillary god Attis. Proper worship of the Great Mother, however, was open only to the wealthy, for salvation could be obtained only through the sacrifice

The Cult of the Great Mother

of the bull. Other long-established deities were the Egyptian gods Isis and Sarapis, whose clergy were particularly honest and fraternal. The cult of Mithra and that of the Syrian gods had expanded more recently, thanks to the perpetual movement of the troops and of the merchants. Mithra, the Sun, representing the supreme god, appealed to the military classes by promising courageous initiates a glorious life in the hereafter. The religion had numerous sanctuaries (*mithraea*) throughout the empire—in the Danubian and Germanic areas, in Britain, and even in Rome. There were large numbers of followers also of the great Syrian Baals: Zeus Dolichenus (from Doliche in northern Syria); Zeus of Heliopolis—Baalbek; and the highland Baal, the El-Gabal (Elagabal) of Emesa. During the Severan era and at their very court, there was developed that syncretism which tended to fuse into a single supreme god all divinities, and especially the solar gods (Sol, Helios, Sarapis, and Mithra). This tendency toward monotheism was accompanied by a lively atmosphere of religiosity and by a taste for magic and miracles.

The rise of Christianity. During the 1st and 2nd centuries, Christianity spread with relative slowness. The doctrines of Jesus, who was crucified about AD 30, first took root among the Jews of Palestine, where a large number of sects were proliferating—orthodox sects, such as the Sadducees and the Pharisees, as well as dissident and sometimes persecuted sects such as the Essenes, whose ascetic practices have been illuminated by the discovery of the Dead Sea Scrolls. At the end of Tiberius' reign, Christianity had spread to the Gentiles as a result of the preaching of St. Paul in Asia Minor and in Greece. At the same time, Christianity continued to make progress among the Jews of Jerusalem, Alexandria, and Syria and quickly reached even Osroëne and the Parthian towns of the Euphrates, where Jewish colonies were numerous. The Roman authorities at first had difficulty in distinguishing the "Christos" believers from the orthodox Jews, but the religion of the former, on leaving its original milieu, quickly became differentiated: to the Romans it appeared illicit because it was mysterious; vile, because its first followers belonged to the lower classes; and dangerous, because, unlike Judaism, which was tolerated as an "ethnic" religion, it was without national ties. The 1st-century historian Tacitus declared that there dwelt in Christians a hatred of humankind. Christians did not participate in the Jewish revolt of 66–73, and under the Flavians, Christianity completely severed itself from its origins. At that time, the East was the centre of the new religion, whose followers were multiplying from Egypt to the Black Sea and were beginning to be noticed in Bithynia and in Greece. Christians seemed fairly numerous in Rome as early as the end of the 1st century. When the age of the Apostles ended, the age of the church began, with its bishops, presbyters, and deacons, with its catechism, preaching, and celebration of the Eucharist. In the 2nd century, Christianity began to reach the intellectuals. Hellenistic culture offered educated Christians the resources of philosophical dialectic and of sophist rhetoric. The example of Philo of Alexandria had shown in the 1st century that it was possible to reconcile the Bible with the great Platonic ideas. By the 2nd century the Christian "apologists" tried to show that Christianity was in harmony with Greco-Roman humanism and that it was intellectually, and above all morally, superior to paganism.

Beginnings of Christian church organization

But the Christians did not succeed in convincing the authorities. The first persecution, that of Nero, was related to a particular incident and had no juridical foundation (the fire at Rome in 64). The concept of Christianity as a crime was undoubtedly established under the Flavians, for about 111, Trajan ordered Pliny, the governor of Bithynia, to put to death all those who declared themselves Christians, without the necessity of their having committed any other crime. At the same time, however, he prescribed that no searches be made for them, and this line of conduct was followed by his successors. Thus, the persecutions remained localized and sporadic and were the result of private denunciations or of spontaneous popular protests. Under Marcus Aurelius, the difficulties of the times often caused the Christians, who refused to sacrifice to the state gods and to participate in the imperial cult, to be accused of provoking the wrath of the gods: martyrs appeared in the East, in Rome, in Gaul, and in Africa. Commodus' reign was more favourable to them, perhaps because certain members of his circle, not a very edifying one in other respects, were Christians or Christian sympathizers. This reprieve, however, was shortlived: Septimius Severus inaugurated the first systematic persecution. In 202 an edict forbade Christian (and Jewish) proselytism. Members of extremist sects were persecuted for preaching continence (which violated Augustus' laws against celibacy), for holding the state in contempt, and especially for refusing military service. Under Caracalla, the situation quieted and the church continued to progress, favoured perhaps by the relative freedom which the law granted to funerary *collegia* (whence the first catacombs).

Cultural life from the Antonines to Constantine. Latin literature enjoyed its "Silver Age" under the Antonines, with the majority of great authors, such as Tacitus, Juvenal, and Pliny the Younger, having begun their careers under Domitian. They had no heirs: after Tacitus, Roman history was reduced to biography. It was only in the 4th century that history began to flourish again, with Ammianus Marcellinus, a Greek writing in Latin. Satire, the Roman genre par excellence, came to an end with Juvenal; and Pliny the Younger, a diligent rhetorician but with a lesser degree of talent, had only the mediocre Fronto as a successor. Another African, Apuleius of Madauros, was more original, being a rhetorician, scholar, and picaresque novelist (*Metamorphoses*).

Decline of Latin literature

A Greek renaissance, however, took place during the 2nd century. The second sophistic school reigned in every area, in rhetoric, history, philosophy, and even in the sciences. Schools of rhetoric and philosophy prospered in the East, in Smyrna, Ephesus, Pergamum, Rhodes, Alexandria, and even in Athens, protected and subsidized by the emperors, from Vespasian to Marcus Aurelius. The great sophists were Herodes Atticus, a multimillionaire from Athens; Polemon; and Aelius Aristides, a valetudinarian devotee of Asclepius. Dio Cassius and Herodian were conscientious and useful historians (first half of the 3rd century), as was later Dexippus the Athenian, whose work survives only in fragments. Science was represented by the mathematician Nicomachus of Gerasa; medicine, by Galen of Pergamum; astronomy, by the Alexandrian Ptolemy. Law remained the only Roman science, exemplified under the Antonines by Salvius Julianus and Gaius (the *Institutiones*) and later rising to its zenith in the 3rd century as a result of the works of three Jurists: Papinian, Ulpian, and Modestinus. Philosophy, heavily influenced by rhetoric and ethics, was represented under Domitian and Trajan by Dio (or Chrysostom) of Prusa, who outlined the stoical doctrine of the ideal sovereign. The biographer Plutarch and Lucian of Samosata were more eclectic, especially Lucian, who resembled Voltaire in his caustic skepticism. Under Marcus Aurelius one of Lucian's friends, Celsus, wrote the first serious criticism of Christianity, "The True Word," known through Origen's refutation of it in the 3rd century. At this time philosophy leaned toward religious mysticism: under the Severans, Ammonius Saccas created the school of Alexandria, and his disciple Plotinus founded the Neoplatonist school, which was to fight bitterly against Christianity. After the apologists and, above all, Tertullian (*c.* 160–after 222), Christian thought deepened, and theology made its appearance. Clement and Origen (*c.* 185–*c.* 254), the greatest theologian of the time, were the luminaries of the church of Alexandria; the Roman church still wrote in Greek and was represented by the slightly old-fashioned Hippolytus; and the church of Africa had a powerful personality, Saint Cyprian, bishop of Carthage.

The Christian theologians

The disappearance of the great lyric and poetic styles, the fossilizing of education as it came to be completely based on rhetoric (*paideia*), and the growing importance of philosophical and religious polemical literature among both pagans and Christians were the basic traits that, as early as the 3rd century, foreshadowed the intellectual life of the late empire.

Military anarchy and the disintegration of the empire (235–270). *Succession of emperors and usurpers.* The

period from the death of Severus Alexander to the time of Claudius II Gothicus was marked by usurpations and barbarian invasions. After Maximinus the Thracian, who bravely fought the Germans but showed great hostility toward the Senate and the educated elite, the Gordians rose to power as a result of a revolt by wealthy African landowners. A senatorial reaction first imposed civilian emperors, Pupienus and Balbinus together, and then named Gordian III, a youth backed by his father-in-law, the praetorian prefect Timesitheus. Gordian III was murdered by the soldiers during a campaign against the Persians and was replaced, first by Philip the Arabian and then by Decius, both soldiers. Decius tried to restore Roman traditions and also persecuted the Christians, but he was killed by the Goths in 251 in a battle near the Black Sea. From 253 to 268 two Roman senators, Valerian and his son Gallienus, reigned. Valerian revived the persecution of the Christians, but he was captured by the Persians during a disastrous campaign and died in captivity (260). His son then reigned alone, facing multiple invasions and several usurpations. He moved constantly between the Rhine and the Danube, achieving brilliant victories (Milan in 262, the Nestus in 267), but the Panonnian army raised several competitors against him (Ingenuus, Regalianus, Aureolus). Too busy to protect the Gauls against the Franks and the Alamanni, and the East against the Persians, he had to tolerate the formation of the Gallic empire under the praetorian prefect Marcus Cassianius Postumus (259–268) and the Palmyrene kingdom of Odaenathus (Odenathus; 260–267). Some of his reforms were a foreshadowing of the future: the senators were practically excluded from the army; the equites received the majority of commands and of provincial governorships; and the composition of the army was modified by the creation of new army corps and especially of a strong cavalry, which was placed under the command of a single leader and charged with closing the breaches that the barbarians were opening along the frontiers. Upon his father's death Gallienus had put an end to the persecution of the Christians, preferring to fight the new religion through intellectual means; to that end, he favoured the ancient Greek cults (Demeter of Eleusis) and protected the Neoplatonist philosopher Plotinus. These initiatives increased the number of his enemies, particularly among the patriotic senators and the Pannonian generals. While Gallienus was in Milan besieging the usurper Aureolus, he was killed by his chiefs of staff, who proclaimed Claudius II (268), the first of the Illyrian emperors. The new emperor won a great victory against the Alamanni on the Garda lake and overwhelmed the Goths in Naissus (269) but died of the plague in 270. This fatal period brought to light one of the major defects of the empire: the lack of a legitimate principle of succession and the preponderant role of the army in politics. The structures that had created the strength of the principate were weakened, and the empire required deep reforms. Gallienus had felt their necessity but had been too weak to impose them.

Victories over the Germans

The barbarian invasions. The Goths were East Germans who came from what is now Sweden and were followed by the Vandals, the Burgundians, and the Gepidae. The after-effect of their march to the southeast, toward the Black Sea, was to push the Marcomanni, the Quadi, and the Sarmatians onto the Roman *limes* in Marcus Aurelius' time. Their presence was brusquely revealed when they attacked the Greek towns on the Black Sea *c.* 238. Timesitheus fought against them under Gordian III, and under Philip and Decius they besieged the towns of Moesia and Thrace, led by their kings, Ostrogotha and Kniva. Beginning in 253, the Crimean Goths and the Heruli appeared and dared to venture on the seas, ravaging the shores of the Black Sea and the Aegean, as well as several Greek towns. In 267 Athens was taken and plundered despite a strong defense by the historian Dexippus. After the victories of Gallienus on the Nestus and Claudius at Naissus (Nish), there was, for a time, less danger. But the countries of the middle Danube were still under pressure by the Marcomanni, Quadi, Iazyges, Sarmatians, and the Carpi of free Dacia, who were later joined by the Roxolani and the Vandals. In spite of stubborn resistance, Dacia was gradually overwhelmed, and it was abandoned by the Roman troops, though not evacuated officially. When Valerian was captured in AD 259/260, the Pannonians were gravely threatened, and Regalianus, one of the usurpers proclaimed by the Pannonian legions, died fighting the invaders. The defense was concentrated around Sirmium and Siscia-Poetovio, the ancient fortresses that had been restored by Gallienus, and many cities were burned.

In the West the invasions were particularly violent. The Germans and the Gauls were driven back several times by the confederated Frankish tribes of the North Sea coast and by the Alamanni from the Middle and Upper Rhine. Gallienus fought bitterly, concentrating his defense around Mainz and Cologne, but the usurpations in Pannonia prevented him from obtaining any lasting results. In 259–260 the Alamanni came through the Agri Decumates (the territory around the Black Forest), which was now lost for the Romans. Some of the Alamanni headed for Italy across the Alpine passes; others attacked Gaul, devastating the entire eastern part of the country. Passing through the Rhône Valley, they eventually reached the Mediterranean; and some bands even continued into Spain. There they joined the Franks, many of whom had come by ship from the North Sea, after having plundered the western part of Gaul. Sailing up the estuaries of the great rivers, they had reached Spain and then, crossing the Strait of Gibraltar, had proceeded to Mauretania Tingitana. Gallienus, outflanked, entrusted Gaul and his young son Saloninus to Postumus, who then killed Saloninus and proclaimed himself emperor. The several invasions had so frightened the people that the new emperor was readily accepted, even in Spain and Britain. He devoted himself first to the defense of the country and was finally considered a legitimate emperor, having established himself as a rival to Gallienus, who had tried in vain to eliminate him but finally had to tolerate him. Postumus governed with moderation, and, in good Roman fashion, minted excellent coins. He, too, was killed by his soldiers, but he had successors who lasted until 274.

Alamanni invasions

Difficulties in the East. In the East the frontiers had been fixed by Hadrian at the Euphrates. But under Nero, the Romans had claimed control over the kings of Armenia, and under Caracalla they had annexed Osroëne and Upper Mesopotamia. The Parthian empire had been weak and often troubled, but the Sāsānids were more dangerous. In 241, Shāpūr I (Sapor), an ambitious organizer and statesman, mounted the throne: he united his empire by bringing the Iranian lords into line and by protecting the Zoroastrian religion. He also tolerated the Manichaeans, put an end to the persecutions of the Christians and Jews, and thereby gained the sympathy of these communities. In 252, with a large army at his command, Shāpūr imposed Artavasdes on Armenia, attacked Mesopotamia, and took Nisibis. In 256 his advance troops entered Cappadocia and Syria and plundered Antioch, while Doura-Europus, on the Middle Euphrates, was likewise falling to him. Valerian had rushed to its aid, but he could not remedy the situation; and in 259 or 260 he was imprisoned by Shāpūr during operations about which little is known. Mesopotamia was lost and Rome was pushed back to the Euphrates. Cappadocia, Cilicia, and Syria were again plundered, and a puppet emperor was appointed in Antioch. But these victories were transitory: in Osroëne, Edessa had shown resistance, a defense was organized in Cappadocia and Cilicia, and Odaenathus, the prince of Palmyra, took Shāpūr by surprise and forced him back to Iran. Having thus aided the Roman cause, Odenathus then began to act in his own interest: he continued the fight against the Persians and took the title "King of Kings." The Romans officially entrusted him with the defense of the East and conferred on him the governorship of several provinces, so that the "kingdom" of Palmyra thus extended from Cilicia to Arabia. He was murdered in 267 without ever having severed his ties with Gallienus. His widow Zenobia had her husband's titles granted to their son Vaballathus. Then in 270, taking advantage of the deaths of Gallienus and Claudius II, she invaded Egypt and a part of Asia Minor with her troops. This invasion was followed by a rupture with Rome, and in 271 Vaballathus was proclaimed

Imperator Caesar Augustus. The latent separatism of the Eastern provinces and, undoubtedly, some commercial advantages caused them to accept Palmyrene domination without difficulty, as they had, in the past, supported Avidius Cassius and Pescennius Niger against the legitimate emperors. In 272 unity was restored by Aurelian, but Mesopotamia was lost and the Euphrates became the new frontier of the empire.

Economic and social crisis. The invasions and the civil wars explain the gravity of the economic and social crisis at the end of Gallienus' reign. Many regions had been laid waste (Gaul, Dacia, Moesia, Thrace, and numerous towns on the Aegean Sea); many important cities had been pillaged or destroyed (Byzantium, Antioch, Olbia, Lyons); and in northern Italy, Gallia Cisalpina (Cisalpine Gaul) itself had been crushed by the Alamanni. The upheavals led to brigandage on land and piracy at sea. Plague, brought back from the East by the armies, ravaged the land for 20 years between 250 and 270. There ensued a population decline, the seriousness of which cannot be estimated; and the empire as a whole was impoverished, although certain provinces, such as Africa, Britain, Pannonia, Asia, and Syria, remained prosperous throughout the century. Municipal government atrophied, inscriptions and dedications became rare, and the cities withdrew into themselves behind narrowed walls. Commerce was often paralyzed by the barbarians; and the main trade route from London to Byzantium, through Boulogne, Cologne, and the Danube, was unsafe and often impassable. The loss of Doura, the defeats inflicted by Shāpūr, and the Palmyrene claims to power slowed down the caravan trade with the East and the Far East. Commerce, however, remained active in the West between Spain, Narbonensis, and Italy, while Africa retained its relations with Ostia, Rome, Egypt, and Syria. Agricultural and handicraft production suffered from insecurity, from military requisitions, and from the abuses of a hard-pressed administration that was attempting by every means to assure the *annona* yield and to provision the large towns so as to prevent outbreaks of social conflict. The empire became a besieged fortress.

Disruption of commerce by the barbarians

The emperors after Commodus—above all Septimius Severus—had been aware of the danger and had adopted such harsh measures as the requisition of material resources and the mobilization of personnel into the service of the state. The municipalities and even the corporations were crushed by the taxes: the collection of the *annona* was a heavy burden for the financially responsible decurions, landholders, and colonists, who were watched over by brutal and sometimes thieving noncommissioned officers. Only a military and bureaucratic monarchy could assure—but at great price certainly—the essential services and obtain an adherence to vital priorities. The early empire's policy of liberalism was at an end; and some modern scholars, such as Rostovtzeff, have spoken of the instituting of "state Socialism"; yet the latter term is fallacious, for there was neither appropriation of the means of production nor redistribution of income to the benefit of the lower classes. Rather, the late empire was a military dictatorship.

Deep social transformations foreshadowed the society of the late empire. The urban middle classes were overburdened, impoverished, and enslaved; and it soon became necessary to halt the flight of the municipal officials by force. The senators lost their political power; but they did preserve their social prestige, and their latifundial fortunes increased at the expense of the decurions and the small landholders. To avoid the obligations of urban life, they tended more and more to retire to their magnificent country villas. The equites thus became the dominant class. They took over the state administration, directed the important offices, controlled finances and taxation, and wielded military and governmental authority in the majority of provinces. Some were civilians, jurists by education; the most powerful came out of the army, where the centurionship had provided open access to high positions since the time of the Severi. During the 3rd century, the army was the empire's rampart against the barbarians, the safest base of imperial power, and the mechanism for social advancement. The lower classes, the humble peasants, the settlers, and the town plebs, officially protected by a state that considered itself egalitarian, suffered from the general impoverishment and the decrease in production: fiscal necessities tended to bind the peasant to his land, the craftsman to his workshop, and the merchant to his trade. Even if the law did not yet prescribe these restrictions, social evolution was moving in this direction, as shown, for example, by the plight of the increasing number of settlers who were forced to cultivate the lands of the rich and of the state.

The recovery of the empire and the establishment of the dominate (270–337). *The Illyrian emperors.* After Claudius II's unexpected death, the empire was ruled from 270 to 284 by several "Illyrian" emperors, who were good generals and who tried in an energetic way to restore equilibrium. The most remarkable was Aurelian. He first gained hard-won victories over the Alamanni and the Juthungi, who had invaded the Alpine provinces and northern Italy. To cheer the inhabitants of Rome, who had succumbed to panic, he began construction of the famous rampart known as Aurelian's Wall. And while crossing the Danubian provinces, before marching against Palmyra, he decided on an orderly evacuation of Dacia, an undefendable region that had been occupied by the barbarians since the time of Gallienus. In the East, he defeated Zenobia's troops easily and occupied Palmyra in 272. Shortly afterward, an uprising broke out in Egypt under the instigation of a rich merchant, who, like a great part of the population, was a partisan of the Palmyrene queen. In response, Aurelian undertook a second campaign, plundering Palmyra and subjugating Alexandria. These troubles, however, along with the devastation of the great caravan city, were to set back Roman trade seriously in the East. Later, rounding back on the Gallic empire of Postumus' successors, he easily defeated Tetricus, a peaceful man and not very willing to fight, near Châlons-sur-Marne. The unity of the empire was restored, and Aurelian celebrated a splendid triumph in Rome. He also reestablished discipline in the state, sternly quelled a riot of artisans in the mints of Rome, organized the provisioning of the city by militarizing several corporations (the bakers, the pork merchants), and tried to stop the inflation by minting an *antoninianus* of sounder value. His religious policy was original: in order to strengthen the moral unity of the empire and his own power, he declared himself to be the protégé of the Sol Invictus (the Invincible Sun) and built a magnificent temple for this god with the Palmyrene spoils. Aurelian was also sometimes officially called *dominus et deus:* the principate had definitely been succeeded by the "dominate." In 275 Aurelian was murdered by certain officers who mistakenly believed that their lives were in danger.

Aurelian's conquests

His successor, for once, was chosen by the Senate—at the army's request and on short notice. But Tacitus, an aged senator, reigned only for a few months. After him, Probus, another Illyrian general, inherited a fortified empire but had to fight hard in Gaul, where serious invasions occurred in 275–277. Afterward, Probus devoted himself to economic restoration; he attempted to return abandoned farmland to cultivation and, with the aid of military labour, undertook works of improvement. To remedy the depopulation, he admitted to the empire, as had Aurelian, a great number of defeated barbarian Goths, Alamanni, and Franks and permitted them to settle on plots of land in Gaul and in the Danubian provinces. After the assassination of Probus in 282 by soldiers, Carus became emperor and immediately associated with himself his two sons, Carinus and Numerian. Carus and Numerian fought a victorious campaign against the Persians but died under unknown circumstances. Carinus, left behind in the West, was later conquered and killed by Diocletian, who was proclaimed emperor in November 284 by the army of the East.

Diocletian. Diocletian may be considered the real founder of the late empire, though the form of government he established—the tetrarchy, or four persons sharing power simultaneously—was transitory. His reforms, however, lasted longer. Military exigencies, and not the desire to apply a preconceived system, explain the successive

nomination of Maximian as Caesar and later as Augustus in 286 and of Constantius and Galerius as Caesars in 293.

The tetrarchy

The tetrarchy was a collegium of emperors comprising two groups: at its head, two Augusti, older men who made the decisions; and in a secondary position, two Caesars, younger, with a more executive role. All four were related either by adoption or by marriage, and all were Illyrians who had attained high commands after a long military career. Of the four, only Diocletian was a statesman. The unity of the empire was safeguarded, despite appearances, for there was no territorial partitioning. Each emperor received troops and a sector of operation: Maximian, Italy and Africa; Constantius, Gaul and Britain; Galerius, the Danubian countries; and Diocletian, the East. Practically all governmental decisions were made by Diocletian, from whom the others had received their power. He legislated, designated consuls, and retained precedence. After 287, he declared his kinship with the god Jupiter (Jove), whom Diocletian claimed was his special protector. Diocletian, together with his Caesar Galerius, formed the "Jovii" dynasty, whereas Maximian and Constantius, claiming descent from the mythical hero Hercules, formed the "Herculii." This "Epiphany of the Tetrarchs" served as the divine foundation of the regime. The ideological recourse to two traditional Roman divinities represented a break with the orientalizing attempts of Elagabalus and Aurelian. Even though he honoured Mithra equally, Diocletian wanted to be seen as continuing the work of Augustus. And in dividing power, Diocletian's aim was to avoid usurpations, or at least to stifle them quickly—as in the attempt of Carausius, chief of the army of Britain, who was killed (293), as was his successor, Allectus (296), after a landing by Constantius.

The deification of the imperial function, marked by elaborate rituals, tended to set the emperors above the rest of mankind. But it was still necessary to avoid future rivalries and to assure the tetrarchy a legitimate and regular succession. Apparently, some time between 300 and 303, Diocletian found an original solution. After the anniversary of their 20-year reign the two Augusti abdicated (Maximian quite unwillingly), and on the same day (May 1, 305) the two Caesars became Augusti. Two new Caesars were chosen, Severus and Maximinus Daia, both friends of Galerius, whose strong personality dominated Constantius. In repudiating the principle of natural heredity (Maximian and Constantius each had an adult son), Diocletian took a great risk: absolute divine monarchy, which Diocletian largely established, implies the hereditary transmission of power, and the future would soon demonstrate the attachment of the troops and even of the population to the hereditary principle.

Changes in provincial organization

In order to create a more efficient unity between subjects and administrators, Diocletian multiplied the number of provinces; even Italy was divided into a dozen small units of the provincial type. This put an end to Italy's privileges, particularly since Rome was no longer the effective capital of the empire, each emperor having his own residence in the part of the empire over which he ruled (Trier, Milan, Sirmium, Nicomedia). Although a few provinces were still governed by senators (proconsuls or consuls), the majority were given to equestrian *praesides,* usually without any military power but responsible for the entirety of civil administration (justice, police, finances, and taxes). The cities lost their autonomy, and the curiales administered and collected the taxes under the governor's direct control. The breaking up of the provinces was compensated for by their regrouping into a dozen dioceses, under equestrian vicars who were responsible to the emperor alone. The two praetorian prefects had less military power but played an important role in legislative, judicial, and above all, financial matters: the administration of the *annona,* which had become the basis of the fiscal system, in fact gave them management of the entire economy. Within the central administration the number of offices increased, their managers being civilians who carried out their functions as a regular career. All officials were enrolled in the *militia,* whose hierarchy was to be outlined during the 4th century.

Great efforts were devoted to strengthening the borders, and the *limes* were outfitted with fortresses (*castella*) and small forts (*burgi*), notably in Syria. The army's strength was increased to 60 legions (but with reduced personnel); and, in principle, each border province received a garrison of two legions complemented by subsidiary troops. Adopting one of Gallienus' ideas, Diocletian created an embryonic tactical army under the direct orders of the emperor whose escort (*comitatus*) it formed. The troops were most often commanded by *duces* and *praepositi,* rather than by provincial governors, and were mainly recruited from among the sons of soldiers and from barbarians who enlisted individually or by whole tribes. In addition, the landowners had to provide either recruits or a corresponding sum of money. All of these reforms were instituted gradually, during defensive wars whose success demonstrated the regime's efficiency. Constantius put down Carausius' attempted usurpation and fought the Alamanni fiercely near Basel; Maximian first hunted down the Bagaudae (gangs of fugitive peasant brigands) in Gaul, then fought the Moorish tribes in Africa, in 296–298, triumphing at Carthage; and on the Danube, Diocletian, and later Galerius, conquered the Bastarnae, the Iazyges, and the Carpi, deporting them in large numbers to the provinces.

Persian opposition

In the East, however, the opposition of the Persians, led by the enterprising Narses, extended from Egypt to Armenia. The Persians incited uprisings by both the Blemmyes nomads in southern Egypt and the Saracens of the Syrian desert, and made use of anti-Roman propaganda by the Manichaeans and Jews. Diocletian succeeded in putting down the revolt in Egypt and fortified the south against the Blemmyes. But in 297, Narses, the heir to Shāpūr's ambitions, precipitated a war by taking Armenia, Osroëne, and part of Syria. After an initial defeat, Galerius won a great victory over Narses, and in 298 the peace of Nisibis reinstated a Roman protégé in Armenia and gave the empire a part of Upper Mesopotamia that extended even beyond the Tigris. Peace was thus assured for some decades.

The wars, the reforms, and the increase in the number of officials were costly, and inflation reduced the resources of the state. The *annona,* set up by Septimius Severus, had proved imperfect, and Diocletian now reformed it through the *jugatio-capitatio* system: henceforth, the land tax, paid in kind by all landowners, would be calculated by the assessment of fiscal units based on extent and quality of land, type of crops grown, number of settlers and cattle, and amount of equipment. The fiscal valuation of each piece of property, estimated in *juga* and *capita* (interchangeable terms whose use varied by region and period of time), required a number of declarations and censuses, similar to those practiced long before in Egypt. Each year, the government established the rate of tax per fiscal unit; and every 15 years, beginning in 312, taxes were reassessed. This complicated system was not carried out uniformly in every region. Nevertheless, it resulted in an improved accounting of the empire's resources and a certain progress in fiscal equity, thus making the administration's heavy demands less unbearable. In addition, Diocletian wished to reorganize the coinage and stabilize inflation. He thus minted improved sterling coins and fixed their value in relation to a gold standard. Nevertheless, inflation again became disturbing by the end of the century, and Diocletian proclaimed his well-known Edictum de Maximis Pretiis, fixing price ceilings for foodstuffs and for goods and services, which could not be exceeded under pain of death. The edict had indifferent results and was scarcely applied, but the inscriptions revealing it have great economic interest.

Social effects of Diocletian's reforms

Diocletian's reforms adumbrated the principal features of late empire society—the lack of social mobility and the omnipotence of the central administration. Fiscal considerations required that the *annona* be collected regularly and that all land be productive. The tendency was to bind the settlers more and more to their land, to prevent villagers from leaving their homes, and to fix everyone to his place of birth. The curiales and their properties were bound to their cities, the corporation members to their trades, and the sons of soldiers were obliged to take their fathers' places in the army. The upper classes were less

tied, having the advantages derived from public careers, whether legally or fraudulently acquired. The senators, now usually called *clarissimi* ("most honorable"), held the major part of landed wealth, and their authority weighed heavily on the peasants. Many of them never attended the Senate in Rome, which had lost its political role. The senatorial class was renewed by the admission of high officials to the clarissimate, including praetorian prefects and numerous provincial governors. The lower appointments were held by the equites, the *perfectissimi* ("the most excellent"), whose place in the social hierarchy depended on their rank in the *militia.* In general, society was becoming hierarchical, and the official who was well paid, both in money and kind, acquired an influence that would only increase during the 4th century.

After a period of initial indifference, Diocletian ended his reign by unleashing against the Christians the last and most violent of their persecutions. Decius and Valerian had already dealt severely with them in the middle of the 3rd century and for the first time had directly struck the church's clergy and property by requiring all inhabitants of the empire to sacrifice to the state gods. There were numerous martyrs, but the persecutions came to a quick end upon the death of Decius in 251 and the capture of Valerian in 260. Gallienus published an edict of tolerance, which enabled the church to resume its progress, thanks to the glory of its martyrs and "confessors," the abundance of its donations (for the rich were now also being reached), and the development of the episcopal organization. At the end of the century the new religion was predominant in certain regions of Asia Minor (Cappadocia), well represented in the East, and even powerful in Rome, where the apostle Peter's successors affirmed their episcopal authority. In Italy, Gaul, and Africa around Carthage, the bishoprics were growing in number, and some councils (or synods) assembled dozens of bishops. The progress of a religion that could not accept the religious basis of the tetrarchy, and certain of whose members were imprudent and provocative, as in the incidents at Nicomedia (where a church was built across from Diocletian's palace), finally aroused Galerius' fanaticism. He, in turn, brought pressure on his Augustus: in 303–304 several edicts, each increasingly stringent, ordered the destruction of the churches, seizure of sacred books, and imprisonment of the clergy, and a sentence of death for all those who refused to sacrifice to the Roman gods. In the East, where Galerius was imposing his ideas more and more on the aging Diocletian, the persecution was extremely violent, especially in Egypt, Palestine, and the Danubian regions. In Italy, Maximian, zealous at the beginning, quickly tired; and in Gaul, Constantius merely destroyed a few churches without carrying reprisals any further. The persecution lasted as late as 312 in the East. Nevertheless, Christianity could no longer be eradicated, for the people of the empire and even some officials no longer felt the blind hatred for Christians that had typified previous centuries.

Struggle for power. The first tetrarchy had ended on May 1, 305; the second did not last long. After Constantius died at Eboracum in 306, the armies of Britain and Gaul, without observing the rules of the tetrarchic system, had hastened to proclaim Constantine, the young son of Constantius, as Augustus. Young Maxentius, the son of Maximian (who had never wanted to retire), thereupon had himself proclaimed in Rome, recalled his father into service, and got rid of Severus. Thus, in 307–308 there was great confusion. Seven emperors had, or pretended to have, the title of Augustus: Maximian, Galerius, Constantine, Maxentius, Maximinus Daia, Licinius (who had been promoted Augustus in 308 by Galerius against Constantine), and, in Africa, the usurper Domitius Alexander also took the purple.

The seven competing emperors

This situation was clarified by successive eliminations. In 310, after numerous intrigues, old Maximian was killed by his son-in-law Constantine, and in the following year Alexander was slain by one of Maxentius' praetorian prefects. In 311 Galerius died of illness a few days after having admitted the failure of his persecutions by proclaiming an edict of tolerance. There remained, in the West, Constantine and Maxentius and in the East, Licinius and Maximinus Daia. Constantine, the best general, invaded Italy with a strong army of faithful Gauls and defeated Maxentius near the Milvian Bridge, not far from Rome. While attempting to escape, Maxentius drowned. Constantine then made an agreement with Licinius, and the two rallied the Eastern Christians to their side by guaranteeing them religious tolerance in the Edict of Milan (313). This left Maximinus Daia, now isolated and regarded as a persecutor, in a weak position; attacked by Licinius near Adrianople, he fell ill and died soon afterward, in 313. This left the empire with two leaders, Constantine and Licinius, allied in outward appearances and now brothers-in-law as a result of Licinius' marriage to Constantine's sister.

The reign of Constantine. Constantine and Licinius soon disputed among themselves for the empire. Constantine attacked his adversary for the first time in 316, taking the dioceses of Pannonia and Moesia from him. A truce between them lasted 10 years. In 316 Diocletian died in Salona, which he had never felt a desire to leave despite the downfall of his political creation. Constantine and Licinius then reverted to the principles of heredity, designating three potential Caesars from among their respective sons, all still infants, with the intention of securing their dynasties (two sons of Constantine and one of Licinius). The dynastic concept, however, required only a single emperor, imposing his own descendance. Although Constantine favoured the Christians, Licinius resumed the persecutions; and in 324 war erupted once again. Licinius, defeated first at Adrianople and then in Asia Minor, was obliged to surrender and, together with his son, was executed. Next, Constantine's third son, Constantius, was in turn named Caesar, as his two elder brothers, Crispus and Constantine the Younger, had been some time before. The second Flavian dynasty was thus founded, and Constantine let it be believed that his father, Flavius Constantius (Chlorus), was descended from Claudius Gothicus.

Constantine's conversion to Christianity had a far-reaching effect. Like his father, he had originally been a votary of the Sun and had gone to worship at the Grand Temple of the Sun in the Vosges Mountains of Gaul, where he had had his first vision—a pagan one. During his campaign against Maxentius, he had had a second vision—a lighted cross in the sky—and he had painted on his men's shields a figure that was perhaps Christ's monogram (although he probably had Christ confused with the Sun in his manifestation as *summa divinitas* ["the highest divinity"]). After his victory he declared himself Christian. His conversion remains somewhat mysterious and his contemporaries—Lactantius and Eusebius of Caesarea—are scarcely enlightening and even rather contradictory on the subject. But it was doubtless a sincere conversion, for Constantine had a religious turn of mind. He was also progressive and greatly influenced by the capable bishops who surrounded him from the very beginning.

Constantine's conversion to Christianity

Until 320–322 solar symbols appeared on his monuments and coins, and he was never a great theologian. But his favourable policy toward the Christians never faltered. Christianity was still a minority religion in the empire, especially in the West and in the countryside (and consequently within his own army), thus excluding the possibility of any political calculation on his part. But it was enthusiastically welcomed in the East, and thanks to Constantine the new religion triumphed more rapidly; his official support led to the conversion of numerous pagans, although with doubtful sincerity because they were indifferent in their moral conviction.

The church, so recently persecuted, was now suddenly showered with favours: the construction of magnificent churches (Rome, Constantinople), donations and grants, exemptions from decurial duties for the clergy, juridical competences for the bishops, and exceptional promotions for Christian officials. Pagans were not persecuted, however, and Constantine retained the title of *pontifex maximus.* But he spoke of the pagan gods with contempt and forbade certain types of worship, principally nocturnal sacrifices. In 331 he ordered an inventory of pagan property, despoiled the temples of their treasure, and finally destroyed a few Eastern sanctuaries on the pretext of immorality.

The churches were soon to feel the burden of imperial solicitude: the "secular arm" was placed at the service of a changing orthodoxy, for the Emperor was impressionable to arguments of various coteries and became quite lost in theological subtleties. In 314 the Council of Arles had tried in vain to stop the Donatist schism (a nationalistic heretical movement questioning the worthiness of certain church officials) that arose in Africa after Diocletian's persecutions. The Arian heresy raised even more difficulties: Arius, an Alexandrian priest and disciple of Lucian of Antioch, questioned the dogma of the Trinity and of the Godhead of Christ, and his asceticism, as well as the sharpness of his dialectics, brought him many followers; he was convicted several times, but the disorders continued. Constantine, solicited by both sides and untroubled by doctrinal nuances that were, moreover, foreign to most Westerners, wished to institute a universal creed; with this in mind he convened the general Council of Nicaea, or Nicene Council, in 325. He condemned Arius and declared, in spite of the Easterners, that Jesus was "of one substance" with God the Father. Nevertheless, the heresy continued to exist, for Constantine changed his mind several times; he was influenced by Arian or semi-Arian bishops and was even baptized on his deathbed, in 337, by one of them, Eusebius of Nicomedia.

The Arian heresy

Between 325 and 337 Constantine effected important reforms, continuing Diocletian's work. The division between the *limitanei* border troops and the tactical troops (*comitatenses* and imperial guard) led by *magistri militum* was clarified, and military careers became independent of civil careers. He conferred the clarissimate upon many equestrian officials, created new titles that recorded personal fidelity (counts and patricians: companions and relatives of the emperor), and gave first rank in the central administration to the palace questor, the *magister officiorum,* and the counts of finance (*comes sacrarum largitionum, comes rei privatae*). The diocesan vicars were made responsible to the praetorian prefects, whose number was increased and whose jurisdictions were now vast territories: the prefectures of Gaul, Italy, Illyricum, and the East. To this unification of political power there corresponded a decentralization of administration.

In order to reorganize finances and currency he minted two new coins: the silver *miliarensis* and, most importantly, the gold *solidus,* whose stability would make it the Byzantine Empire's basic currency. And by plundering Licinius' treasury and despoiling the pagan temples, he was able to restore the finances of the state. Even so, he still had to create class-taxes: the *gleba* for senators, and the *chrysargyre,* which was levied in gold and silver on merchants and craftsmen in the towns.

Constantine's immortality, however, rests on his founding of Constantinople. This "New Rome," established in 324 on the site of Byzantium and dedicated in 330, rapidly increased in population as a result of favours granted to immigrants. A large number of churches were also built there, even though former temples were not destroyed; and the city became the administrative capital of the empire, receiving a senate and proconsul. This choice of site was due not to religious considerations, as has been suggested, but rather to reasons that were both strategic (its proximity to the Danube and Euphrates frontiers) and economic (the importance of the straits and of the junction between the great continental road, which went from Boulogne to the Black Sea, and the eastern commercial routes, passing through Asia Minor to Antioch and Alexandria). Constantine died on May 22, 337.

Founding of Constantinople

The Roman Empire under the 4th-century successors of Constantine. *The rule of Constantine's sons.* After some months of confusion, Constantine's three surviving sons (Crispus, the eldest son, had been executed in mysterious circumstances in 326), supported by the armies faithful to their father's memory, divided the empire among themselves and had all the other members of their family killed. Constantine II kept the West, Constantius the East, and Constans, the youngest brother, received the central prefecture (Italy, Africa, and Illyricum). In 340, Constantine II tried to take this away from Constans but was killed. For the next 10 years there was peace between the two remaining brothers, and Constans won acceptance for a religious policy favourable to the Niceans, whose leader, Athanasius, had received a triumph in Alexandria. In 350 a mutiny broke out in Autun; Constans fled but was killed in Lyons by Magnentius, a usurper who was recognized in Gaul, Africa, and Italy. Constantius went out to engage Magnentius, and the Battle of Mursa (351) left the two strongest armies of the empire—those of Gaul and of the Danube—massacred, thus compromising the empire's defense. Magnentius retreated after his defeat and finally committed suicide in 353.

Thenceforth, Constantius reigned alone as Augustus, aided by a meddlesome bureaucracy in which mission deputies (*agentes in rebus*), informers, and spies played an important role. He named two Caesars in succession, his two young surviving cousins, Gallus in the East and Julian in Gaul, both of whom were apparently considered only "employees," or *apparitores.* Constantius eventually had to get rid of Gallus, who proved incompetent and cruel and soon terrorized Antioch. Julian, however, was a magnificient success, a fact that aroused Constantius' jealousy and led to Julian's usurpation; for the latter was proclaimed Augustus, in spite of Constantius' opposition, at Lutetia in 361. Civil war was averted when Constantius died in November 361, leaving the empire to Julian, the last ruler of the Constantinian family.

The reign of Constantius

At the time of his death in 337, Constantine had been preparing to go to war against the Persians. This legacy weighed heavily on the shoulders of Constantius, a military incompetent when compared to the energetic Sāsānian king Shāpūr II. Nearly every year the Persians attacked and pillaged Roman territory; the Mesopotamian towns were besieged, and Nisibis alone resisted. There was a lull between 350 and 357, while Shāpūr was detained by troubles in the eastern regions of his own kingdom. The war resumed, however, and Mesopotamia was partly lost when the emperor had to go off to fight Julian. Constantius had fought Shāpūr conscientiously, but his generals were mediocre, except for Urisicinus, and he himself was clumsy. In the meantime, the Rhine and the Danube, the troops of which had been withdrawn in favour of the East, were threatened frequently; and Constantius had made a mistake in sending Chnodomar, the Alamannic king, against Magnentius in 351, for his tribes had gone on to ravage Gaul. Julian, however, soon revealed himself to be a great military leader by winning several well-fought campaigns between 356 and 361, most notably at Strasbourg in 357, and by restoring approximately 70 plundered villages. Constantius defeated the Quadi and the Goths on the Danube in 359, but court intrigues, Magnentius' usurpation, and the interminable war against the Persians allowed the barbarians to wreak great havoc.

Constantius was primarily interested in religious affairs, and his interventions created a "caesaro-papism" that was unfavourable to the church, for after the Battle of Mursa the Emperor had become violently Arian. The Christological problem had moved to the forefront. In 360 Constantius obtained a new creed by force from the Council of Constantinople, which, rejecting the notion of "substance" as too risky, declared only that the Son was like the Father and thus left the problem unresolved. Pagans as well as orthodox Nicaeans (Homoousians) and extremist Arians (Anomoeans) were persecuted, for in 356–357 several edicts proscribed magic, divination, and sacrifices and ordered that the temples be closed. But when Constantius visited Rome in 357, he was so struck by its pagan grandeur that he apparently suspended the application of these measures.

Constantius' Arianism

The reign of Julian. Julian, who had been spared because of his tender age from the family butchering in 337, had been brought up far from the court and was undoubtedly intended for the priesthood. Nevertheless, he had been allowed to take courses in rhetoric and philosophy at Ephesus and later at Athens; he developed a fondness for Hellenic literature, and he secretly apostatized around 351. When he became sole emperor at the end of 361, he proclaimed his pagan faith, ordered the restitution of the temples seized under Constantius, and freed all the bishops who had been banished by the Arians, so as to

weaken Christianity through the resumption of doctrinal disputes. His paganism was irrational, mystical, and complex. Privately, he worshipped the Sun (Helios-King) and the figure of Cybele, the pagan equivalent of the Virgin Mary. Deeply attached to the traditional Hellenism of the Sophists, who ceaselessly invoked Homer and Plato, he was especially influenced by the last of the Neoplatonists, Iamblichus and Maximus of Ephesus, who instructed him in the techniques of ecstasy and direct communication with the gods (theurgy). The gods spoke to him, and he heard their exhortations. Julian sincerely believed that it was his mission to reestablish their worship.

In time Julian would undoubtedly have become a persecutor, for in 362 he had already forbidden Christians from teaching, under the pretext that their faith made them betray the intentions of the ancient authors. This ostracism aroused even the indignation of certain tolerant pagans. His restoration of the pagan temples provoked disorders and judicial violations among Christians. He gave the pagans a monopoly on high posts, and as a consequence, numerous self-seeking Christians became pagans. Within a short while Julian was successful enough in his undertaking to have aroused the fear and hatred of the Christians, who for a long time thought of him as the Antichrist.

In the political realm, Julian wished to return to the liberal principate of the Antonines—to a time before the reforms of Diocletian and Constantine, whom he detested. He put an end to the terrorism of Constantius' eunuchs and *agentes in rebus* and reduced the personnel and expenditures of the court, while he himself lived like an ascetic. In the provinces he lightened the financial burden on individuals by reducing the *capitatio*, and on cities, by reducing the *aurum coronarium* and restoring the municipal properties confiscated by Constantius. On the other hand, he increased the number of curiales by reinstating numerous clerks in an attempt to return the ancient lustre to municipal life. Thus, he earned the gratitude of pagan intellectuals, who were enamoured of the past of free Greece; and Ammianus made him the central hero of his history.

Julian's attempt to defeat Persia

Taking up Trajan's dream, Julian wished to defeat Persia definitively by engaging the empire's forces in an offensive war that would facilitate a national reconciliation around the gods of paganism. But his army was weak—corrupted perhaps by large numbers of hostile Christians. After a brilliant beginning, he was defeated near Ctesiphon and had to retrace his steps painfully; he was killed in an obscure encounter on June 26, 363.

Julian's successor, Jovian, chosen by the army's general staff, was a Christian, but not a fanatic. He negotiated a peace with Shāpūr, by which Rome lost a good part of Galerian's conquests of 298 (including Nisibis, which had not surrendered) and abandoned Armenia. He also restored tolerance in religious affairs, for he neither espoused any of the heresies nor persecuted pagans. In February 364 he died accidentally.

The reign of Valentinian and Valens. Once again the general staff unanimously chose a Pannonian officer—Valentinian, an energetic patriot and, like Jovian, a moderate Christian—but he had to yield to the rivalry of the armies by dividing authority. Taking the West for himself, Valentinian entrusted the East to his brother Valens, an inexperienced man whom he raised to the rank of Augustus. For the first time the two parts of the empire were truly separate, except for the selection of consuls, in which Valentinian had precedence.

Although he served the state with dedication, Valentinian could be brutal, choleric, and authoritarian. His foreign policy was excellent: all the while he was fighting barbarians (the Alemanni in Gaul, the Sarmatians and Quadi in Pannonia) and putting down revolts in Britain and Africa (notably that of the Berber Firmus) with the aid of his top general, Theodosius the Elder, he was taking care to improve the army's equipment and to protect Gaul by creating a brilliant fortification. His domestic measures favoured the curiales and the lower classes: from then on, taxes would be collected exclusively by officials; the protection of the poor was entrusted to "defenders of the plebs," chosen from among retired high officials (*honorati*). Nevertheless, needs of state obliged him to accentuate social immobility, to reinforce corporation discipline and official hierarchization, and to demand taxes ruthlessly. At first he was benevolent to the Senate of Rome, supervised the provisioning of the city, and legislated in favour of its university, the nursery of officials (law of 370). But beginning in 369, under the influence of Maximin, the prefect of Gaul, he initiated a period of terror, which struck the great senatorial families. Meanwhile, religious peace reigned in the West, tolerance was proclaimed, and after some difficulty, Rome found a great pope in Damasus, who, beginning in 373, actively supported the new bishop of Milan, St. Ambrose, an ardent defender of orthodoxy.

Religious peace in the West, persecutions in the East

In the East, Valens, who was incapable and suspicious, had fallen under the influence of legists, such as the praetorian prefect Modestus. The beginning of Valens' reign was shadowed by the attempted usurpation of Procopius (365–366), a pagan relative of Julian's who failed and was killed by the army, which remained faithful to Valens. Modestus instituted harsh persecutions in Antioch of the educated pagan elite. Valens was a fanatic Arian, who exiled even moderate Nicaean bishops and granted to Arians favours that aroused violent reactions from the orthodox, whose power had increased in the East. Valens' policies made the East prey to violent religious passions.

On the Danube, Valens fought the Visigoths and made a treaty with their king, Athanaric, in 369; but in 375 the Ostrogoths and the Greutingi appeared on the frontiers, pushed from their home in southern Russia by the powerful Huns. In 376 Valens authorized the starving masses to enter Thrace; but being exploited and mistreated by the officials, they soon turned to uncontrollable pillaging. Their numbers continually increased by the addition of new bands, until finally they threatened Constantinople itself. Valens sent for aid from the West; but without waiting for it to arrive, he joined battle and was killed in the Adrianople disaster of 378, which to some critics foreshadowed the approaching fall of the Roman Empire.

The Goths, who were also stirring up Thrace and Macedonia, could no longer be driven out. The provinces remained insecure until 382, when Theodosius, after fighting many hard battles with only an inconclusive result, resigned himself to settling the Goths legally in the Balkan Peninsula as *foederati* (soldiers in the service of Rome, receiving subsidies and furnishing recruits); there they formed barbarian enclaves.

The reign of Gratian and Theodosius I. Following Valentinian's sudden death in 375, the West was governed by his son Gratian, then 16 years old, who had been given the title of Augustus as early as 367. The Pannonian army, rife with intrigue, quickly proclaimed Gratian's half-brother, Valentinian II, only four years old. The latter received Illyricum under his older brother's guardianship, and this arrangement satisfied everybody. Valentinian's advisers were executed: Maximian was sacrificed to the spite of the Senate, and Theodosius the Elder became the victim of personal jealousies. Gratian announced a liberal principate, supported in Gaul by the wealthy family of the Bordeaux poet Ausonius and in Rome by the Symmachi and the Nicomachi Flaviani, representatives of the pagan aristocracy. His generals defeated the Alamanni and the Goths on the Danube but arrived too late to save Valens.

On January 19, 379, before the army, Gratian proclaimed Theodosius, the son of the recently executed general, as Eastern emperor. Theodosius was chosen for his military ability and for his orthodoxy (Gratian, extremely pious, had come under the influence of Damasus and Ambrose). The East was enlarged by the dioceses of Dacia and Macedonia, taken from Valentinian II. Gratian and Theodosius agreed to admit the Goths into the empire, and Gratian applied the policy also to the Salian Franks in Germany. Theodosius soon dominated his weak colleague and entered the battle for the triumph of orthodoxy. In 380 the Arians were relieved of their churches in Constantinople, and in 381, the Nicaean faith was universally imposed by a council whose canons established the authority of the metroplitan bishops over their dioceses and gave the bishop of the capital a primacy similar to that of the bishop of Rome.

Influx of Goths and Franks

In ecclesiastical affairs, the separation between East and West was codified. The Westerners bowed to this policy, satisfied with the triumph of orthodoxy. Gratian then permitted Ambrose and Damasus to deal harshly with the Arians, with the support of the state. Paganism also was hounded: following Theodosius' lead, Gratian refused the chief priesthood, removed the altar of Victory from the hall of the Roman Senate, and deprived the pagan priests and the vestals of their subsidies and privileges. The pagan senators were outraged, but their protests were futile because Gratian was watched over by Ambrose.

This militantly orthodox policy aroused the displeasure of the pagans and of the Western Arians: thus, when Gratian left Trier for Milan, the army of Gaul and Britain proclaimed its leader, Maximus, in 383. He conquered Gaul without difficulty, and Gratian was killed in Lyons. Maximus, who, like Theodosius, was Spanish and extremely orthodox, was recognized by the latter. In the meantime, the third Augustus, Valentinian II, had taken refuge in Milan after suffering defeat in Pannonia. He was effectively under the domination of his mother, Justina, an Arian who sought support for her son among the Arians and pagans of Rome and even among the African Donatists. In 388 Maximus, after arriving in Italy, first expelled Valentinian and then prepared to attack Theodosius. The latter, accepting the inevitability of war, strengthened his resolve and gained several victories. Maximus was killed at Aquileia in 388, and thenceforth Theodosius ruled both West and East; he was represented in the East by his son Arcadius, an Augustus since 383. Valentinian II was sent to Trier, accompanied by the Frankish general Arbogast to control him.

Sole rulership of Theodosius

After a few years' respite, during the prefectureships of Nicomachus Flavianus in Rome and Tatian in the East, paganism waged its last fight: Theodosius, influenced by Ambrose, who had dared to inflict public penance on him in 390 after the massacre at Thessalonica, had determined to eliminate the pagans completely. After a few hostile clashes, the law of November 8, 392, proscribed the pagan religion. Then Arbogast, after Valentinian II's death in 392 under shadowy circumstances, proclaimed as emperor the rhetorician Eugenius. When Theodosius refused to recognize him, Eugenius was thrown into the arms of the pagans of Rome. But this last "pagan reaction" was short-lived; in 394, with his victory at the Frigidus River, between Aquileia and Emona, Theodosius put an end to the hopes of Eugenius and his followers. His intention was to place his son Honorius, proclaimed Augustus in 393, over the West, while returning his eldest son, Arcadius, to the East. But Theodosius' sudden death, in January 395, precipitated the division of the empire.

Theodosius had finally eliminated the danger of the Goths, although not without taking risks, and had both established a dynasty and imposed the strictest orthodoxy. A compromise peace with the Persians had given Rome, in 387, a small section of Armenia, where he had founded Theodosiopolis (Erzurum). Internally, there were few reforms; and wars, wastefulness, and the burden of administration had caused a noticeable increase in taxation. The ancient authors all paint a gloomy picture of the condition of the settlers, the curiales, and the artisans of the towns. Meanwhile, the laws were becoming increasingly harsh, and revolts (Antioch in 387, Thessalonica in 390) were ruthlessly suppressed. Theodosius owed his success to the support of the orthodox bishops—who took advantage of his faith to introduce "clericalism" into the state—and to the strength of his army, hereafter dominated by barbarian elements: the Franks and Alemanni in the West; the Goths and even the federated Huns in the East. The majority of commanding officers were also barbarians, including such Franks as Richomer, Merovech, Arbogast, and the half-Vandal Stilicho, who through his marriage to Serena, Theodosius' niece, had entered the imperial family.

Social and economic conditions. During the 4th century the emperor's power was theoretically absolute, the traditions of the principate having given way to the necessities of defense.

The emperor, who was most often a soldier, was both heir to the Hellenistic *basileus* (absolute king) and the anointed of the deity. Pagans and Christians alike considered him "emperor by the grace of God," which, strictly speaking, rendered the imperial cult unnecessary. Indeed, only military victory, willed by God, distinguished the legitimate sovereign from the "tyrant." After Constantine, administration was decentralized and headed by praetorian prefects with enormous power. A gargantuan and strictly hierarchic bureaucracy restricted direct action by the emperor, who was separated increasingly from his subjects by a hierarchical court, particularly under Constantius and Theodosius. This overrun administration was recruited from among students of literature and of law for higher posts, and for lesser posts from among "notaries," who knew how to write shorthand. Under Constantius the latter attained even the highest positions, in spite of their modest origins. There was superabundant legislation as well, which tended to increase fiscal income by limiting exemptions and privileges and to keep each person in the rank assigned him at birth, for taxes were based on social classes. Tax abuses provoked uprisings by the lower classes as well as recourse to patronage and attempts to escape one's class by entering the administration and, above all, to get into the clarissimate, which was lucrative and rich in privileges.

Hierarchized bureaucracy

Apart from the tiny elite of Roman senators, for whom public office was only a last resort, and the great land holders sheltered in their villas, where in the West at least they were able to live in independence, officials everywhere formed the upper class of society. The equites occupied the inferior and beginning positions, for gradually during the course of the century all the high officials attained the clarissimate. A career in public service could no longer be compared to the old magistracies. Public service was a powerful factor in social advancement (like the army in the 3rd century), but it was also a costly and sometimes corrupt superstructure, whose upkeep drained the empire. To protect themselves, the poor and weak took refuge in the patronage (*patrocinium*) of the powerful (*potentes*), a method which, while it shielded them from the authority of the state in spite of the laws, also placed them in positions of personal dependence, which foreshadowed the social arrangements of the Middle Ages. The bishops, who held spiritual authority over growing portions of the population and had available to them both the power of their office and large revenues, began to concentrate the defense of local interests in their own hands and to compete with the rich for the protection of the weak. In the 5th century, they would be the masters and protectors of the cities, particularly in the West.

Economic life was based on ownership of land, and the number of free peasants and middle-sized landowners was decreasing; for both groups were burdened by heavy taxes and threatened by the expansionism of the rich estate (*latifundium*) owners, the senators and high functionaries. Some free villages (*metrocomiae*) continued to exist, notably in Syria, and were protected by the state. In areas where urbanization was sparse, the *latifundium* dominated and was worked by settlers. The *adscripticius* settler was subordinated completely to his master (*dominus*) and owed him forced labour, whose name (*obsequia servilia*) showed its true nature. The *tributarius* settler was a freeholder but still had to look to the patronage of the powerful man whose land he cultivated. The laws specified that the settler could not leave his land, and a recaptured fugitive was treated like a slave, even though he remained a free man in principle. This evolution toward serfdom was ancient; but it had been accelerated by the fiscal reforms (the *jugatio,* the taxes in kind, and the increases in amounts of forced labour due). In sum, the settler system was one aspect of the general social immobility of the late empire, and it resulted from the enormous needs of the state.

Settler system

The economy was of three types, which were unequally distributed: 1) the closed economy of the large estates, which dominated in the few regions where urbanization was weak; 2) the produce state economy, in which taxes, wages, and salaries were paid in kind so as to reduce the disadvantages of inflation for the treasury—but the possibility, often resorted to, of replacing payment in kind with payment in money (*adaeratio*) probably allowed the state

to speculate on monetary fluctuations and to satisfy one or the other social category; 3) the monetary economy, the most widespread, which had been restored by Constantine: the gold standard (the *solidus,* or gold pound) was used to establish prices, to pay certain taxes and services, and to assist in the amassing of wealth (coins, bars, ingots); the regular currency, the *nummus,* was a silver-covered copper coin, having various names and values.

Although there was considerable inflation (culminating under Theodosius), in spite of a deflationary fiscal policy, commercial transactions ignored barter and were based instead on currency throughout the empire at the end of the century. The economy was partially under state direction, which was applied to agriculture through bias toward the settler system on imperial estates and to industry through the requisitioning of corporations (artisans, merchants, carriers) and the creation of state workshops (especially for manufacturing military goods). Opinions differ on the intensity of trade, but there was certainly clear progress in comparison to the 3rd century.

Comparison between East and West

It is above all necessary to distinguish between the East and West. The West was in decline: the rural areas had taken over from the cities, where activity had become anemic; and autonomous blocs had formed in Gaul (with the relative prosperity of Britain), Africa, Italy (which suffered a marked decline, except in Cisalpina), and the Danubian countries. The great West–East route of the early empire (from Boulogne to Byzantium, by way of the Rhine and the Danube), as a result of threats by barbarians, had been displaced toward the south, through Lyons, Milan, and Aquileia. On the other hand, the East was less affected by invasions; there, urban life was traditionally more active, and the peasantry resisted the settler system and patronage more effectively. Although the Balkans were devastated

From W. Shepherd, *Historical Atlas*, Harper & Row, Publishers (Barnes & Noble Books), New York; revision Copyright © 1964 by Barnes & Noble, Inc.



The Roman Empire in the late 4th century AD.

after Adrianople (378), elsewhere the wealth of the towns and of the officials sustained active trading. There were three great commercial arteries: 1) the "wheat" road, from Alexandria to Constantinople by way of Antioch; 2) the Far Eastern trade routes, through Osroëne (Batnae) or Cappadocia (Somoata) to Antioch and the capital; and 3) the maritime route linking East to West, from Seleucia (the port of Antioch) to Rome. In addition, there were the more dangerous land routes, such as the one across the Balkans (the ancient Via Egnatia through Macedonia) or the Danube route (Sardica, Sirmium, Aquincum, Carnuntum). Although Egypt was impoverished from having been exploited for so long, Asia Minor around Constantinople was prosperous, having numerous towns as well as well-populated rural areas. And of the Eastern provinces, Syria was the richest—thanks to a balanced agriculture, a large and stable population, the industries of Antioch, the textiles of Apamea, Damascus, and Berytus, and above all thanks to commerce, which was centred in Antioch and its port, Seleucia, rebuilt by Constantius. All of this accounts for the intensity of urbanization.

The remnants of pagan culture. The spread of Christianity in no way harmed the flourishing of pagan literature. Instruction in the universities (Rome, Milan, Carthage, Bordeaux, Athens, Constantinople, Antioch, Alexandria) was still based on rhetoric, and literature received the support of senatorial circles, especially in Rome (for example those of the Symmachi and the Nicomachi Flaviani). Latin literature was represented by Symmachus and the poet Ausonius. The last great historian of Rome was Ammianus Marcellinus, a Greek who wrote in Latin for the Roman aristocracy; of his *Res gestate,* the most completely preserved part describes the period from 353 to 378. The works of the abridgers Aurelius Victor and Eutropius are fairly accurate and more reliable than the *Scriptores historiae Augustae,* a collection of imperial biographies of unequal value, undoubtedly composed under Theodosius but for an unknown purpose. Erudition was greatly prized in aristocratic circles, which, enamoured of the past, studied and commented on the classic authors (Virgil) or the ancestral rites (the *Saturnalia* of Macrobius). Greek literature is represented by the works of philosophers or sophists: Themistius, a political theoretician who advocated absolutism; Himerius of Prusias; and above all Libanius of Antioch, whose correspondence and political discourses from the Theodosian period bears witness to his perspicacity and, often, to his courage.

The Christian Church. From the time of Constantine the church had perfected its organization, modelled on that of the state. Each city had its bishop; each provincial capital had its Metropolitan, and some had a recognized "primacy": the patriarchs of Alexandria, Antioch, and Constantinople and the bishops of Milan and Rome, the latter claiming an *auctoritas* over the others, which was a heritage of classic Rome. The church had become extremely wealthy and had obtained immunities for its clergy and its bishops. The emperors had assumed the right to interfere in religious affairs, but under Theodosius, Pope Damasus and St. Ambrose reacted: the state was to restrict itself to furnishing the "secular arm," while the church, in the name of evangelical ethics, claimed the right to judge the emperors, a policy that had grave implications for the future. The "caesaro-papism" of Constantius would gain adherents under the Byzantine emperors. Paganism, officially suppressed by Theodosius, survived well into the 5th century in the rural areas, where missionary activity (such as that of St. Martin in Gaul) was only recent, and among the intellectuals, teachers, and senators of Rome.

Triumph of orthodoxy

Orthodoxy triumphed over the great heresies (Donatism, Arianism), but other heresies continued to emerge. In the meantime, the Goths had been converted to Arianism by Ulfilas during the period of Constantius and Valens, thus presaging conflicts that would come after the great invasions. Orthodox missionaries had converted Osroëne, Armenia, and even some countries on the Red Sea. And in Egypt at the end of the 3rd century there appeared the monasticism of isolated hermits (St. Anthony) and of cenobites grouped in communities (Pachomius). Monasticism rapidly spread to Syria and Asia Minor by toning down its ascetic severity, and in the West St. Martin founded the monasteries of Ligugé and Marmoutier.

The Christian literature of the 4th century is remarkable. Its first representative is Eusebius of Caesarea, a friend and panegyrist of Constantine and a church historian whose creation of a "political theology" sealed the union between the Christian emperor and the church. St. Athanasius wrote apologetic works and a life of St. Anthony. Also prominent were the great Cappadocians: St. Basil of Caesarea, St. Gregory of Nazianzus, St. Gregory of Nyssa, and St. John Chrysostom of Antioch, the greatest preacher of his time. The Westerners, too, had great scholars and brilliant writers: St. Hilary of Poitiers, enemy of the Arians and of Constantius; St. Ambrose, administrator and pastor, whose excessive authority was imposed on Gratian and even on Theodosius; and St. Jerome, a desert monk and confessor of upper class Roman ladies, a formidable polemicist who knew Greek and Hebrew and made the first faithful translation of the Old and New Testaments (the Vulgate) as well as of a chronicle of world history, which was a translation and continuation of the work of Eusebius. Finally, St. Augustine, the bishop of Hippo, was a great pastor, a vigorous controversialist, a sensitive and passionate writer (the *Confessions*), and the powerful theologian of *The City of God.* The century that developed these great minds cannot be considered decadent.

The eclipse of the Roman Empire in the West (c. 395–500) and the German migrations. *Invasions in the early 5th century.* After the death of Theodosius the Western empire was governed by young Honorius. Stilicho, an experienced statesman and general, was charged with assisting him and maintaining unity with the East, which had been entrusted to Arcadius. The Eastern leaders soon rejected Stilicho's tutelage. An anti-barbarian reaction had developed in Constantinople, thus impeding the objectives of the half-Vandal Stilicho. He wanted to intervene on several occasions but was prevented from doing so by a threat from the Visigoth chieftain Alaric, whom he checked at Pollentia in 402, then by the Ostrogoth Radagaisus' raid in 406, and finally by the great invasion of the Gauls in 407. The following year he hoped to restore unity by installing a new emperor in Constantinople, Theodosius II, the son of Arcadius, who had died prematurely; but he succumbed to a political and military plot in August 408. The division of the two *partes imperii* was now a permanent one.

Honorius, seated in Ravenna, a city easier to defend than Milan, had only incompetent courtiers surrounding him, themselves animated by a violent hatred of the barbarians. Alaric soon reappeared, at the head of his Visigoths, demanding land and money. Tired of the Romans' double-dealing, he descended on Rome itself. The city was taken and pillaged for three days, thus putting an end to an era of Western history (August 410). An Arian, Alaric spared the churches. He died shortly therafter in the south; and his successor, Athaulf, left the peninsula to march against the Gauls.

Sack of Rome by the Visigoths

Fleeing from the terrifying advance of the Huns, on December 31, 406, the Vandals, Suebi, and Alani, immediately followed by the Burgundians and bands of Alemanni, crossed the frozen Rhine and swept through Gaul, effortlessly throwing back the federated Franks and Alemanni from the frontiers. Between 409 and 415 a great many of these barbarians arrived in Spain and settled in Lusitania (Suebi) and in Baetica (Vandals, whence the name Andalusia). As soon as Gaul had become slightly more peaceful, Athaulf's Visigoths arrived, establishing themselves in Narbonensis and Aquitania. After recognizing them as "federates," Honorius asked them to go to Spain to fight the Vandals. Meanwhile, the Roman general Constantius eliminated several usurpers in Gaul, confined the Goths in Aquitania, and reorganized the administration (the Gallic assembly of 418). But he was unable to expel the Franks, the Alemanni, and the Burgundians, who had occupied the northern part of the country, nor to eliminate the brigandage of the Bagaudae. He was associated with the empire and was proclaimed Augustus in 421, but he died shortly afterward. His son, Valentinian III, succeeded Honorius in 423 and reigned until 455.

The beginning of Germanic hegemony in the West. During the first half of the 5th century the barbarians gradually installed themselves, in spite of the efforts of the Roman general Aetius at the head of a small army of mercenaries and of Huns. Aetius took back Arles and Narbonne from the Visigoths in 436, either pushed back the Salian and Ripuarian Franks beyond the Rhine or incorporated them as federates, settled the defeated Burgundians in Sapaudia (Savoy), and established the Alani in Orléans. The other provinces were lost: Britain, having been abandoned in 407 and already invaded by the Picts and Scots, fell to the Angles, Saxons, and Jutes; a great Suebi kingdom, officially federated but in fact independent, was organized in Spain after the departure of the Vandals, and it allied itself to the Visigoths of Theodoric I, who were settled in the country around the Garonne.

Loss of provinces

In 428 the Vandal Gaiseric led his people (80,000 persons, including 15,000 warriors) to Africa. St. Augustine died in 430 in besieged Hippo, Carthage fell in 435, and in 442 a treaty gave Gaiseric the rich provinces of Byzacena and Numidia. From there he was able to starve Rome, threaten Sicily, and close off the western basin of the Mediterranean to the Byzantines.

Shortly afterward, in 450, Attila's Huns invaded the West—first Gaul, where, after having been kept out of Paris, they were defeated by Aetius on the Campus Mauriacus (near Troyes), then Italy, which they evacuated soon after having received tribute from the pope, St. Leo. Attila died shortly afterward; and this invasion, which indeed left more legendary memories than actual ruins, had shown that a solidarity had been created between the Gallo-Romans and their barbarian occupiers, for the Franks, the Alemanni, and even Theodoric's Visigoths had come to Aetius' aid.

After the death of Aetius, in 454, and of Valentinian III, in 455, the West became the stake in the intrigues of the German chiefs Ricimer, Orestes, and Odoacer, who maintained real power through puppet emperors. In 457–461 the energetic Majorian reestablished imperial authority in southern Gaul until he was defeated by Gaiseric and assassinated shortly afterward. Finally, in 476, Odoacer deposed the last emperor, Romulus Augustulus, had himself proclaimed king in the barbaric fashion, and governed Italy with moderation under the theoretical tutelage of the emperor of the East. The end of the Roman Empire of the West passed almost unperceived.

The end of the Roman Empire of the West

Barbarian kingdoms. Several barbarian kingdoms were then set up: in Africa, Gaiseric's Vandals; in Spain and in Gaul as far as the Loire, the Visigothic kingdom; and farther to the north, the Salian Franks and the Alemanni. The barbarians were everywhere a small minority. They established themselves on the great estates and divided the land to the benefit of the federates without doing much harm to the lower classes or disturbing the economy. The old inhabitants lived under Roman law, while the barbarians kept their own "personality of laws," of which the most well-known is the judicial composition, the Wergild. Romans and barbarians coexisted but uneasily. Among the obstacles to reconciliation were differences in mores; social and political institutions (personal monarchies, fidelity of man to man); language, although Latin was still used in administration; and above all religion: the Arianism of the barbarians permitted the Catholic bishops to retain their hold over their flocks. The only persecution, however, was under the Vandals, whose domination was the most harsh.

Two great kingdoms marked the end of the 5th century. In Gaul, Clovis, the king of the Salian Franks (reigned 481/482–511), expelled Syagrius, the last Roman, from Soissons, took Alsace and the Palatinate from the Alamanni (496), and killed Alaric II, king of the Visigoths, at Vouillé (507). His apparently sudden conversion to Catholicism assured him the support of the bishops, and Frankish domination was established in Gaul. At the same time, Theodoric, king of the Ostrogoths, was reigning in Italy. He had been charged by the emperor Zeno to take back Italy from Odoacer in 488, and in 494 he had himself proclaimed king at Ravenna. His Goths, few in number, were established in the north; elsewhere he preserved the old imperial administration, with senators as prefects. Externally, he kept Clovis from reaching the Mediterranean and extended his state up to the valley of the Rhône. Theodoric died in 526. Ten years later Justinian charged his general Belisarius with the reconquest of Italy, a costly, devastating, and temporary operation that lasted from 535 to 540.

Analysis of the decline and fall. The causes of the fall of the empire have aroused a good deal of discussion. The reasons most often invoked are not definitive; the economic decline of the Roman world was not obvious in the 4th century, nor was the decay of the form of government or the defects of the administration. Nevertheless, the empire was yielding under the weight of its state superstructure and of taxation that was too heavy for the ancient world, where production and the standard of living were both very low. The army, which was extremely expensive, was not large enough; and even the population seemed to have declined. But it was really the massive invasions, which had become incessant during the second half of the century, that toppled the empire. It is essential to see why the East resisted, whereas the West was overcome. The East during this period was rich and well populated, and its heavy state structure was healthier than that of the West. The latter was in the hands of the great landholders, who exploited their settlers, who, in turn, had no recourse. The East was also less threatened from the outside and easier to defend: only the lower Danube was vulnerable, with Persia in fact representing no great danger, whereas in the West long frontiers required large numbers of troops and ruinously expensive fortifications. Once the borders were broken through, the vital regions of Gaul and Italy were left defenseless. Even if the East had come to its aid instead of growing increasingly aloof after the failure of Stilicho, the barbarians would undoubtedly still have triumphed; for the population, although it dreaded the barbarians and hated their religion, showed a surprising apathy: bureaucratic absolutism and the patronage of the powerful in the West had killed the idea of the state and the patriotism of ancient Rome.

(P.P.)

BIBLIOGRAPHY

Aegean civilizations: For recent work see the accounts that appear each year in *The American Journal of Archaeology, Archaeological Reports* (published by the Society for the Promotion of Hellenic Studies and the British School of Archaeology at Athens), *Bulletin de Correspondance Hellénique,* and Greek periodicals like *Arkhaiologikon Deltion. Cambridge Ancient History,* 3rd ed., vol. 1–2 (1970–72); REYNOLD A. HIGGINS, *Minoan and Mycenaean Art* (1967); SINCLAIR HOOD, *The Home of the Heroes: The Aegean Before the Greeks* (1967); SPYRIDON MARINATOS and MAX HIRMER, *Krîti kai Mikinaikí Hellás* (1959; Eng. trans., *Crete and Mycenae,* 1960); FRIEDRICH MATZ, *Kreta und frühes Griechenland* (1962; Eng. trans., *Crete and Early Greece,* 1963). For the Aegean Bronze Age as a background to Homer, see the sections in *Archaeologia Homerica* by various authors since 1967. H.L. LORIMER, *Homer and the Monuments* (1950), is still valuable. For the end of the Bronze Age, see VINCENT DESBOROUGH, *The Last Mycenaeans and Their Successors* (1964). (*Crete*): A.J. EVANS, *The Palace of Minos . . . at Knossos,* 4 vol. (1921–35, reprinted 1964), is still basic; SINCLAIR HOOD, *The Minoans: Crete in the Bronze Age* (1971); CHRISTIAN ZERVOS, *L'Art de la Crète néolithique et minoenne* (1956), for pictures. (*The Cyclades*): CHRISTIAN ZERVOS, *L'Art des Cyclades* (1957), for pictures and references. (*The mainland*): G.E. MYLONAS, *Mycenae and the Mycenaean Age* (1966), for the late Bronze Age and Mycenaean shaft grave circle B; EMILY T. VERMEULE, *Greece in the Bronze Age* (1964), is the standard work. (*Architecture*): J.W. GRAHAM, *The Palaces of Crete* (1962). (*Religion*): M.P. NILSSON, *The Minoan-Mycenaean Religion and Its Survival in Greek Religion,* 2nd ed. (1950), remains the standard work. (*Seals*): JOHN BOARDMAN, *Greek Gems and Finger Rings* (1970). (*Writing and the decipherment of linear B*): JOHN CHADWICK, *The Decipherment of Linear B* (1958); MICHAEL VENTRIS and JOHN CHADWICK, *Documents in Mycenaean Greek* (1956), with a good survey of Aegean writing. References for authorities against the decipherment are listed by Hood in *The Minoans,* p. 167 (*op. cit.*).

Ancient Greek civilization: For the period from *c.* 1200–500 BC, ANTONY ANDREWES, *The Greeks* (1967), presents the historian's view; C.M. BOWRA, *The Greek Experience* (1957), approaches the Greeks through their literature. N.G.L. HAMMOND,

A History of Greece to 322 B.C., 2nd ed. (1967), gives a full account and is mainly devoted to political history. See also *The Cambridge Ancient History*, vol. 3–7 (1925–28), and some relevant chapters in the new edition of vol. 2, which is appearing in fascicles (1970–). MICHAEL GRANT (ed.), *The Birth of Western Civilisation: Greece and Rome* (1964), contains essays by different scholars on Greek history, art, and thought (fully illustrated). (*Geography of the Greek homeland*): MAX CARY, *Geographic Background of Greek and Roman History* (1949), is a useful treatment; and J.L. MYRES, *Geographical History in Greek Lands* (1953), contains essays including studies of geography and colonization. (*Sources for the study of Greek civilization*): MICHAEL GRANT, *The Ancient Historians* (1969), is of much value. P.L. MACKENDRICK, *The Greek Stones Speak* (1962), contains a survey by periods and sites of archaeological discoveries in Greece. The principal ancient historical sources, all available in translation, are HERODOTUS (Archaic period), THUCYDIDES (Peloponnesian War), DIODORUS (5th–4th century), ARRIAN (Alexander), XENOPHON (4th century and the East), PAUSANIAS (description of Greece), and PLUTARCH (biographies). (*The Early Archaic period*): V.R.d'A. DESBOROUGH, *The Last Mycenaeans and Their Successors* (1964) and *Protogeometric Pottery* (1952), contain detailed studies of the full range of archaeological evidence. J.N. COLDSTREAM, *Greek Geometric Pottery* (1968), includes a survey of the archaeological evidence for the history of the period. A.M. SNODGRASS, *The Dark Age of Greece* (1971), provides a full archaeological synthesis of the evidence. C.G. STARR, *The Origins of Greek Civilization, 1100–650 BC* (1961), gives a historian's view of the period *c.* 1200 to 500 BC, M.I. FINLEY, *The World of Odysseus*, rev. ed. (1962), projects Homer's world into the Dark Ages. (*The Greek colonizing movement*): JOHN BOARDMAN, *The Greeks Overseas* (1974), provides a survey of the archaeology of colonizing and trading posts. T.J. DUNBABIN, *The Greeks and Their Eastern Neighbours* (1957), contains lectures on relations and trade with the East. M.M. AUSTIN, *Greece and Egypt in the Archaic Age* (1970), is a short but detailed study of mercenaries and trade in Egypt. T.J. DUNBABIN, *The Western Greeks* (1948), gives a full survey of all sources for south Italy and Sicily. (*Archaic Greek culture*): G.M.A. RICHTER, *Archaic Greek Art Against Its Historical Background* (1949), is a survey by regions and periods; and JOHN BOARDMAN, *Pre-Classical: Style and Civilization* (1967), a brief study of art and society in the Archaic period. A.R. BURN, *The Lyric Age of Greece* (1960), is a good study of all sources with accounts of literature, art, and thought, as well as political and social history. See also W.K.C. GUTHRIE, *A History of Greek Philosophy*, vol. 1, *The Earlier Presocratics and the Pythagoreans* (1962); G.E.R. LLOYD, *Early Greek Science: Thales to Aristotle* (1970); and KARL SCHEFOLD, *Frühgriechische Sagenbilder* (1964; Eng. trans., *Myth and Legend in Early Greek Art*, 1966).

For the period of the development of the Greek *poleis* (*c.* 750–*c.*500 BC), HERODOTUS' *History* is the most valuable source. ARISTOTLE's *Constitution of the Athenians* is useful, especially on Solon and Cleisthenes. PLUTARCH's lives of Lycurgus, Theseus, and Solon preserve some valuable facts. Fragments of the poems of Tyrtaeus and Solon with translations are accessible in J.M. EDMONDS, *Elegy and Iambus*, vol. 1 ("Loeb Classical Library," 1931). Twentieth-century scholarship begins with K.J. BELOCH, *Griechische Geschichte*, 2nd ed., 4 vol. in 8 (1912–27); the first volume deals with the period before the Persian Wars. J.B. BURY, *A History of Greece*, 3rd ed. rev. by RUSSELL MEIGGS (1951), has still to be surpassed as a one-volume survey. ANTONY ANDREWES, *The Greeks* (1967), offers a good account, more descriptive than narrative. W.G. FORREST, *The Emergence of Greek Democracy, 800–400 B.C.* (1966), provides a brief and thoughtful introduction. *The Cambridge Ancient History*, vol. 3–4 (1929–39), though old, is still useful, especially on interstate relations. (*Tyranny and hoplite tactics*): ANTONY ANDREWES, *The Greek Tyrants* (1956), collects together plentiful information. H.T. WADE-GERY, "The Growth of the Dorian States," ch. 22 of *The Cambridge Ancient History*, vol. 3, presents admirably one theory of tyranny. A.M. SNODGRASS, "The Hoplite Reform and History," *Journal of Hellenic Studies*, 85:110–122 (1965), supersedes previous work on the beginnings of hoplite tactics. The same author's *Arms and Armour of the Greeks* (1967) gives a more discursive account of hoplite tactics. (*Sparta*): H.T. WADE-GERY, "The Spartan Rhetra in Plutarch, *Lycurgus VI*," *Classical Quarterly*, 37:62–72 (1943) and 38:1–9, 115–126 (1944), reprinted in *Essays in Greek History* (1958), is a good starting point for the study of early Sparta. ANTONY ANDREWES, *Probouleusis, Sparta's Contribution to the Technique of Government* (1954), offers a theory of conflict between an aristocracy and a wider class. G.L. HUXLEY, *Early Sparta* (1962), is provocative. W.G. FORREST, *A History of Sparta 950–192 BC* (1968), is a brief survey. ANTONY ANDREWES, "The Government of Classical Sparta," in E. BADIAN (ed.), *Ancient Society and Institutions: Studies Presented to Victor Ehrenberg on His 75th Birthday* (1966), makes important points. (*Athens*): CHARLES HIGNETT, *A History of the Athenian Constitution to the End of the Fifth Century B.C.* (1952), offers a view on most issues. H.T. WADE-GERY, *Essays in Greek History* (1958), includes several articles on constitutional questions. B.L. BAILEY, "The Export of Attic Black-Figure Ware," *Journal of Hellenic Studies*, 60:60–70 (1940), has fundamental importance for the economic history of Archaic Athens. ALFRED FRENCH, *The Growth of the Athenian Economy* (1964), is a good treatment of a difficult subject. Two valuable studies of the sources for Athenian constitutional history are: FELIX JACOBY, *Atthis, the Local Chronicles of Ancient Athens* (1949); and JAMES H. DAY and MORTIMER CHAMBERS, *Aristotle's History of Athenian Democracy* (1962). Two specialized studies have brought considerable advance in understanding Dracon and Solon: EBERHARD RUSCHENBUSCH, "Φονος zum Recht Drakons und seiner Bedeutung für das Werden des athenischen Staates," *Historia*, 9: 129–154 (1960) and *Σολωνος νομοι: Die Fragmente des solonischen Gesetzeswerkes* (1966).

The main sources for the period *c.* 500 to 404 BC are HERODOTUS (trans. by G. RAWLINSON, introd. by F.R.B. GODOLPHIN, 1942) and THUCYDIDES (trans. by R. CRAWLEY, introd. by J.H. FINLEY, 1950). Herodotus records in detail the Persian invasion of Greece (Books VII–IX), with a long introduction on the events that led up to it. He had to rely largely on oral tradition and prejudiced sources and is unreliable on strategy and tactics, but his narrative is vivid, colourful, and very readable. Thucydides, writing a generation later, is a more mature and critical historian. His theme is the Peloponnesian War, with a brief introduction on the period between the two great wars. His main interests are the nature of power and the reaction of people to events, but he attached great importance to accuracy in the detail of his narrative. These literary sources are supplemented by inscriptions recording decrees, dedications, financial accounts, and casualty lists, especially from Athens; texts of the most significant, with commentary by RUSSELL MEIGGS and DAVID LEWIS, are in *A Selection of Greek Historical Inscriptions to the End of the Fifth Century B.C.* (1969). PLUTARCH, writing some 500 years later, in his lives of Aristides, Themistocles, Cimon, Pericles, Nicias, Alcibiades, and Lysander (*Plutarch's Lives*, trans. by JOHN DRYDEN), preserves important material from earlier sources that are lost, though he is primarily interested in the characters of his heroes and is reluctant to dismiss a good story. There are two good recent studies of Xerxes' invasion: A.R. BURN, *Persia and the Greeks* (1962); and CHARLES HIGNETT, *Xerxes' Invasion of Greece* (1963). DONALD KAGAN, *The Outbreak of the Peloponnesian War* (1969), discusses the policies that led up to the war, which are differently interpreted by G.E.M. DE ST. CROIX, *The Origins of the Peloponnesian War* (1972). RUSSELL MEIGGS, *The Athenian Empire* (1972), analyzes the evidence for the growth and character of the Athenian Empire. Of general histories of Greece, GEORGE GROTE, *History of Greece*, new ed. (1888, reprinted 1971), is still worth reading; N.G.L. HAMMOND, *A History of Greece to 322 B.C.*, 2nd ed. (1967), is detailed and up-to-date, but compressed; VICTOR EHRENBERG, *From Solon to Socrates* (1967), is also comprehensive and easier to read. At shorter length, J.B. BURY, *A History of Greece to the Death of Alexander the Great*, 4th ed. (1974), has been revised to take account of new evidence. For the economic background HUMFREY MICHELL, *The Economics of Ancient Greece*, 2nd ed. (1957), is a useful review of the evidence; ALFRED FRENCH, *The Growth of the Athenian Economy* (1964), studies Athens in more detail. CHARLES HIGNETT, *A History of the Athenian Constitution* (1952), is a good critical review, often unorthodox, of Athenian politics and constitutional changes. W.G. FORREST, *The Emergence of Greek Democracy, 800–400 B.C.* (1966), is the most refreshing new look at Athenian politics and can be usefully supplemented by the examination of the mechanics of Athenian democracy in A.H.M. JONES, *Athenian Democracy* (1966), and the detailed analysis in P.J. RHODES, *The Athenian Boule* (1972). The best general account of Greek living conditions is ROBERT FLACELIERE, *La Vie quotidienne en Grèce au siècle de Périclès* (1959; Eng. trans., *Daily Life in Greece at the Time of Pericles*, 1960); the Athenian scene is gaily illustrated in the comedies of Aristophanes, which combine an excellent sense of theatre with strong political satire, fantasy, and boisterous bawdry.

A select bibliography of the history of the Greek world in the 4th century must necessarily concentrate on general works, which themselves contain more detailed bibliographies. The basic works for the history of 4th-century Greece in French, English, and German, are: GUSTAVE GLOTZ and ROBERT COHEN, *Histoire grecque*, vol. 3, *La Grèce au IVe siècle* (1936) and vol. 4 (pt. 1), with the assistance of PIERRE ROUSSEL, *Alexandre et le démembrement de son empire*, 2nd ed. (1945); *The Cambridge Ancient History*, vol. 4, 3rd ed., *Macedon (401–301 BC)* (1953), better informed than the preceding on the most recent works; HERMANN BENGTSON, *Griechische Geschichte*, 3rd ed. (1965). ROBERT COHEN, *La Grèce et l'hellénisation du monde antique*,

new ed. (1948), is rich in ideas and precious for its indication of the sources and the present state of the questions. For the literary history of the period, still of use is the old and monumental *Histoire de la littérature grecque,* by the brothers ALFRED and MAURICE CROISET, 4th ed., vol. 4 (1928), but ALBIN LESKY carries forward the essential material in his volume of the *Geschichte der griechischen Literatur,* 2nd ed. (1963; Eng. trans., *A History of Greek Literature,* 1966), and his bibliography is more important, more precise, and more up-to-date than that of the Croisets. The diplomatic history is studied more particularly, and with much justice, by the Swiss Hellenist VICTOR MARTIN, *La Vie internationale dans la Grèce des cités, VIe–IVe s. av. J.-C.* (1940). On the civilization in the largest sense of the term, see the two volumes, whose text and illustrations are equally remarkable, of PIERRE LEVEQUE, *L'Aventure grecque* (1964); and of FRANCOIS CHAMOUX, *La Civilisation grecque* (1963); although dealing with nearly identical subjects, the two books, without supplanting one another, complete one another very well. The following are more detailed or more specialized works: The history of Demosthenes and of Philip is explained with finesse and clarity, despite the complications, by PAUL CLOCHE in *Un Fondateur d'Empire: Philippe II, roi de Macédoine* (1955); and WERNER JAEGER, *Demosthenes: The Origin and Growth of His Policy* (1938). The problems of the education of the era are studied thoroughly by H.I. MARROU, *Histoire de l'éducation dans l'antiquité,* 5th ed. (1960; Eng. trans., *A History of Education in Antiquity,* 1956, reprinted 1964), but not without some prejudices. In the matter of art, out of a superabundant bibliography three works stand out: one old but not out-of-date book that explains the evolution of the art of Greece in all of its forms in relation to its history, ANDRE DE RIDDER and W. DEONNA, *L'Art en Grèce* (1924; Eng. trans., *Art in Greece,* 1927, reprinted 1968); the great work of CHARLES PICARD, *Manuel d'archéologie grecque, la sculpture,* vol. 3–4 (1948–63); and a survey work—intelligent, well informed, and well illustrated—by FRANCOIS CHAMOUX, *Art grec* (1966). On Plato the book that can give the best general idea about his policy and philosophy is that of the great specialist AUGUSTE DIES, *Autour de Platon,* 2 vol. (1927), to which it is necessary to add the more summary *Platon,* by the same author (1930). EDOUARD DELEBECQUE, *Essai sur la vie de Xénophon* (1957), studies the relationships between this disciple of Socrates and his times, and attempts to fix the chronology. The best general work on Alexander is GEORGES RADET, *Alexandre le Grand* (1931, reprinted 1950); the bibliography had been brought up-to-date by W. SESTON.

Good sources for the Hellenistic Age (323–30 BC) include *The Cambridge Ancient History,* vol. 7–9 (1928–32), by various authors, with good bibliography; MAX CARY, *A History of the Greek World from 323 to 146 B.C.,* 2nd ed. rev. (1951); W.W. TARN and G.T. GRIFFITH, *Hellenistic Civilisation,* 3rd rev. ed. (1952, reprinted 1966), a good conspectus; PIERRE JOUGUET, *L'Impérialisme macédonien et l'hellénisation de l'Orient* (1926; Eng. trans., *Macedonian Imperialism and the Hellenization of the East,* 1928). An excellent and more recent account of the political history, with full references, is EDOUARD WILL, *Histoire politique du monde hellénistique (323–30 av. J.-C.)* (1966–67). (*Hellenistic kingdoms—Egypt*): E.R. BEVAN, *A History of Egypt Under the Ptolemaic Dynasty* (1927); P.M. FRASER, *Ptolemaic Alexandria* (1972). (*Seleucid Asia*): E.R. BEVAN, *The House of Seleucus,* 2 vol. (1902, reprinted 1966); ELIAS BICKERMAN, *Institutions des Séleucides* (1938). (*Bactria and India*): W.W. TARN, *The Greeks in Bactria and India,* 2nd ed. (1951), fundamental if rather imaginative; A.K. NARAIN, *The Indo-Greeks* (1957). (*Pergamum*): E.V. HANSEN, *The Attalids of Pergamon,* rev. 2nd ed. (1971); R.B. MCSHANE, *The Foreign Policy of the Attalids of Pergamon* (1964). (*Macedonia*): W.W. TARN, *Antigonos Gonatas* (1913); F.W. WALBANK, *Philip V of Macedon* (1940, reprinted 1967). (*Greek states*): J.A.O. LARSEN, *Greek Federal States* (1968); W.S. FERGUSON, *Hellenistic Athens* (1911); HUMFREY MICHELL, *Sparta* (1952). (*Hellenistic politics, institutions, and society—monarchy*): LUCIEN CERFAUX and J. TONDRIAU, *Le Culte des souverains dans la civilisation gréco-romaine* (1957). (*Organization*): HERMANN BENGTSON, *Die Strategie in der hellenistischen Zeit,* 3 vol. (1937–52). (*Hellenization*): MOSES HADAS, *Hellenistic Culture: Fusion and Diffusion* (1959). (*Economy*): M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World,* 2nd rev. ed. (1957). (*Cultural activities*): MARGARETE BIEBER, *The Sculpture of the Hellenistic Age,* rev. ed. (1961); T.B.L. WEBSTER, *Hellenistic Poetry and Art* (1964), and *Hellenistic Art* (1967); R.E. WYCHERLEY, *How the Greeks Built Cities,* 2nd ed. (1962). (*Warfare*): W.W. TARN, *Hellenistic Military and Naval Developments* (1930); E.W. MARSDEN, *Greek and Roman Artillery: Historical Development* (1969), and *Greek and Roman Artillery: Technical Treatises* (1971). (*View of world*): J.O. THOMSON, *History of Ancient Geography* (1948). (*Roman period*): HERMANN BENGTSON, *Griechische Geschichte von den Anfängen bis in die römische Kaiserzeit* (1969).

Ancient Italic peoples: Basic, though partly outdated, is J. WHATMOUGH, *The Foundations of Roman Italy* (1937). See also D. RANDALL-MACIVER, *Italy Before the Romans* (1928); M. PALLOTTINO, "Popolazioni storiche dell'Italia antica," *Guida allo studio della civiltà romana antica,* vol. 1, pp. 71–90 (1952), and *Civilità artistica etrusco-italica,* pp. 9–39 (1971); and G.M.A. RICHTER, *Ancient Italy* (1955). The chief problems are discussed in the light of the post-World War II discoveries by J. HEURGON in *Rome et la Méditerranée occidentale jusqu'aux guerres puniques* (1969). For the languages, see E. PULGRAM, *The Tongues of Italy: Prehistory and History* (1958); V. PISANI, *Le lingue dell'Italia antica oltre il latino,* 2nd ed. (1964). Leading papers on this subject matter are published in the *Studi Etruschi* (annually since 1927). An exhaustive summary of present knowledge of the Etruscan world is offered by M. PALLOTTINO, *Etruscologia,* 6th ed. (1968; Eng. trans., *The Etruscans,* 1956). See also D. STRONG, *The Early Etruscans* (1968). Among the most valuable postwar works dealing with other individual Italic peoples, civilization, and languages are: L. BERNABO BREA, *Sicily Before the Greeks* (1957); R. PERONI, *Archeologia della Puglia preistorica* (1967); O. PARLANGELI, *Studi messapici* (1960); A. ALFOLDI, *Early Rome and the Latins* (1965); G. GIACOMELLI, *La lingua falisca* (1963); G. DEVOTO, *Gli antichi Italici,* 3rd ed. (1967); E.T. SALMON, *Samnium and the Samnites* (1967); G.A. MANSUELLI and R. SCARANI, *L'Emilia prima dei Romani* (1961); and G.B. PELLEGRINI and A.L. PROSDOCIMI, *La lingua venetica,* 2 vol. (1967).

Ancient Rome: Works on Rome from its origins through the years of the early Republic (to 264 BC) include P.G. GIEROW, *The Iron Age Culture of Latium:* vol. 1, *Classification and Analysis* and vol. 2, *Excavations and Finds* (1964), a systematic view of the archaeological discoveries made in Latium, dating from the early Iron Age; EINAR GJERSTAD, *Early Rome:* vol. 1, *Stratigraphical Researches in the Forum Romanum and Along the Sacra Via* (1953); vol. 2, *The Tombs* (1956); vol. 3, *Fortifications, Domestic Architecture, Sanctuaries, Stratigraphical Excavations* (1960); vol. 4, *Synthesis of Archaeological Evidence* (1966), a monumental work on all the archaeological remains from ancient Rome—the investigation of the material is outstanding, although the historical conclusions have been contested; PIETRO DE FRANCISCI, *Primordia Civitatis* (1959), and RAYMOND BLOCH, *Les Origines de Rome,* rev. ed. (1959; Eng. trans., *The Origins of Rome,* 1960), two overall views of the early centuries of Rome internally and externally; ANDRAS ALFOLDI, *Early Rome and the Latins* (1965), a first-rate study of the relations between Rome and the Latins, up to the time of Rome's subjugation of Latium; T.R.S. BROUGHTON, *The Magistrates of the Roman Republic,* 2 vol., and one suppl. vol. (1951–52), an excellent view of the Roman magistracies of the republican era; PIERRE LEVEQUE, *Pyrrhos* (1957), a monograph (in French) that deals with the personality and ambitions of Pyrrhus, Rome's formidable opponent at the beginning of the 3rd century BC; RUDI THOMSEN, *Early Roman Coinage,* 3 vol. (1957–61), a balanced, well-done history of the oldest Roman currency.

Useful works on the middle years of the Republic (264–133 BC) include important sections, together with inscriptions, published in English translation by NAPHTALI LEWIS and MEYER REINHOLD (eds.), *Roman Civilization Sourcebook,* vol. 1, *The Republic* (1966); and A.H.M. JONES, *A History of Rome Through the Fifth Century,* vol. 1, *The Republic* (1968). (*General histories of the period*): *The Cambridge Ancient History,* vol. 7–8 (1928–30); G. DE SANCTIS, *Storia dei Romani,* 2nd ed., vol. 3–4 (1967–69); H.H. SCULLARD, *A History of the Roman World from 753 to 146 B.C.,* 3rd ed. (1969). (*Roman constitutional history*): A.H.J. GREENIDGE, *Roman Public Life* (1901); W.M.F. JASHEMSKI, *The Origins and History of the Pro-Consular and Pro-Praetorian Imperium to 27 B.C.* (1950); THEODOR MOMMSEN, *Römische Staatsrecht,* 3 vol. (1871–88, reprinted 1952); E.T. SALMON, *Roman Colonization Under the Republic* (1969); L. ROSS TAYLOR, *Roman Voting Assemblies* (1966) and *The Voting Districts of the Roman Republic* (1960); A.N. SHERWIN WHITE, *The Roman Citizenship* (1939). (*Social and economic history*): TENNEY FRANK, *An Economic History of Rome to the End of the Republic,* 2nd ed. (1927), and (ed.), *An Economic Survey of Ancient Rome,* 6 vol. (1933–40); M.I. ROSTOVTZEFF, *The Social and Economic History of the Hellenistic World* (1941); ARNOLD TOYNBEE, *Hannibal's Legacy,* 2 vol. (1965). (*Provincial government and imperialism*): E. BADIAN, *Foreign Clientelae, 264–70 B.C.* (1958); TENNEY FRANK, *Roman Imperialism* (1914); DAVID MAGIE, *Roman Rule in Asia Minor to the End of the Third Century After Christ,* 2 vol. (1950); G.H. STEVENSON, *Roman Provincial Administration till the Age of the Antonines* (1939); C.H.V. SUTHERLAND, *The Romans in Spain, 217 B.C.–A.D. 177* (1939).

The two main surviving authorities on the Punic Wars are Polybius and Livy, namely Polybius bk. 1 (First War), bks. 2 and 3, and fragmentary accounts in books 7–15 (Second War),

and fragments of bks. 36–38 (Third War); and Livy bks. 21–30 (Second War), based partly on Polybius and partly on less reliable Roman annalists. Apart from Appian's account of the Third War (*Libyca,* 67–135, based partly upon Polybius), the subsidiary authors add little of value. Modern works on the Punic Wars are found in TENNEY FRANK *et al., Cambridge Ancient History,* vol. 7, ch. 21, and vol. 8, ch. 2–5, 15 (1928–30); H.H. SCULLARD, *A History of the Roman World, 753 to 146 B.C.,* 3rd ed. (1961), and ARNOLD J. TOYNBEE, *Hannibal's Legacy,* 2 vols. (1965). An indispensable work is F.W. WALBANK, *Historical Commentary on Polybius,* vol. 1 (1957), vol. 2 (1967), for full discussion of all problems. G. DE SANCTIS, *Storia dei Romani,* vol. 3, pt. 1–2 (1916–17), and vol. 4, pt. 3 (1964), is fundamental. U. KAHRSTEDT, *Geschichte der Karthager von 218–146* (1913), provides detailed source-criticism. J. KROMAYER and G. VEITH, *Antike Schlachtfelder,* vol. 3–4 (1912–31), and *Schlachten-Atlas zur antiken Kriegsgeschichte,* Röm. Abt. 1–2 (1922), are the standard works on military aspects, while sea-power is handled by J.H. THIEL, *History of Roman Sea-Power Before the Second Punic War* (1954), and *Studies on the History of Roman Sea-Power in Republican Times* (1946), ch. 2 for the Second Punic War. S. GSELL, *Histoire ancienne de l'Afrique du Nord,* vol. 1–4, esp. vol. 3 (1913–20), deals fully with the Carthaginian side. On the Spanish and African campaigns, see H.H. SCULLARD, *Scipio Africanus in the Second Punic War* (1930), and *Scipio Africanus: Soldier and Politician* (1970). For the Third Punic War, see A.E. ASTIN, *Scipio Aemilianus* (1967). On Hannibal, see E. GROAG, *Hannibal als Politiker* (1929).

The chief ancient sources for the late Republic (133–31 BC) are the works of CAESAR and CICERO and SALLUST's two monographs on Jugurtha and Catiline; the biographies of prominent Romans of the period in PLUTARCH, *Parallel Lives;* and the historical narratives of APPIAN (*The Civil Wars*) and DIO CASSIUS (*Roman History*). All are available in the original, with facing English translation, in the "Loeb Classical Library," the first four in many other translations. An important source collection is A.H.J. GREENIDGE and A.M. CLAY, *Sources for Roman History, 133–70 B.C.,* 2nd ed. rev. by E.W. GRAY (1960). The best outline in English is the first half of H.H. SCULLARD, *From the Gracchi to Nero,* 3rd ed. (1970), with excellent notes and bibliography. The classic reference work is WILHELM DRUMANN, *Geschichte Roms in seinem Übergange von der republikanischen zur monarchischen Verfassung,* 2nd ed. by P. GROEBE, 6 vol. (1899), giving biographies (with full source material) of all prominent figures of the period, arranged by families. For a survey of scholarship, with critical bibliography, see E. BADIAN, "From the Gracchi to Sulla (1940–59)," *Historia,* 11:197–245 (1962); part 2 of Badian's *Foreign Clientelae, 264–70 B.C.* (1958), treats the political history down to 70 BC. For the social background, see the fundamental work of MATTHIAS GELZER, *Die Nobilität der römischen Republik* (1912; Eng. trans., *The Roman Nobility,* 1969); and, on special aspects of it, SUSAN TREGGIARI, *Roman Freedmen During the Late Republic* (1969); T.P. WISEMAN, *New Men in the Roman Senate, 139 B.C.–A.D. 14* (1971); and E. BADIAN, *Publicans and Sinners* (1972). L. ROSS TAYLOR, *Party Politics in the Age of Caesar* (1949), gives the best analysis of the working of the political system after Sulla; RONALD SYME, *The Roman Revolution* (1939), is the classic study of the social and political transformation of the republic into the principate. Special aspects of particular interest are studied in JACQUES HARMAND, *L'Armée et le soldat á Rome de 107 à 50 avant notre Ére* (1967); E. BADIAN, *Roman Imperialism in the Late Republic,* 2nd ed. (1968); E.S. GRUEN, *Roman Politics and the Criminal Courts, 149–78 B.C.* (1968); and the last chapters of E.T. SALMON, *Roman Colonization Under the Republic* (1969).

TACITUS' *Annals* and *Histories,* describing the period from AD 14 to 96, are largely lost, including his account of the Flavian period, much of which he himself witnessed. The extant portions, however, written from the viewpoint of a disillusioned member of the Roman Senate, form the authoritative basis for any account of Tiberius, Claudius, Nero, and the year 69. The surviving ancient text that comes closest to covering the whole period of the early Roman Empire (31 BC–AD 193) is DIO CASSIUS' history of Rome to AD 229, in Greek. It is extant for the years 31 BC to AD 46 (and, for those from 46 to 193, there are epitomes and fragments, some of considerable length). These historical works are supplemented by imperial biographies: SUETONIUS' *Lives of the Caesars* (Julius Caesar to Domitian inclusive), PLUTARCH's *Galba* and *Otho,* and the *Historia Augusta* (Hadrian and his successors). Suetonius is very informative and the *Historia Augusta,* a late compilation of dubious authorship and varying value, is a main source for the 2nd century. A selective, but very good collection of source material for the period in English is provided by NAPHTALI LEWIS and MEYER REINHOLD (eds.), *Roman Civilization,* 2 vol. (1951–55). (*Modern works*): *The Cambridge Ancient History,* vol. 10–11 (1934–36), with serviceable bibliographies of earlier work, is still a most helpful study in English. The best one-volume general accounts, both recent and with first-class bibliographies, are ALBINO GARZETTI, *L'Impero da Tiberio agli Antonini* (1960), good for political and social developments; and PAUL PETIT, *La Paix romaine* (1967), good on military and economic aspects. For the foundation of the principate, see D.C. EARL, *The Age of Augustus* (1968), a realistic account; and for the development of the principate, MASON HAMMOND, *The Antonine Monarchy* (1959), careful, but somewhat optimistic. For imperial biographies, see under the individual emperors. ANTHONY BIRLEY, *Marcus Aurelius* (1966); and B.H. WARMINGTON, *Nero* (1969), are very informative, the latter on the first century and the former on the second. For the economy: TENNEY FRANK (ed.), *An Economic Survey of Ancient Rome,* 6 vol. (1933–40)—vol. 2–5 are fundamental for this period; M.I. ROSTOVTZEFF, *The Social and Economic History of the Roman Empire,* 2nd ed. rev. by P.M. FRASER, 2 vol. (1957), brilliant, but often subjective. For social life: JEAN GAGE, *Les Classes sociales dans l'Empire romain* (1964), comprehensive. For the provinces: THEODOR MOMMSEN, *Römische Geschichte,* vol. 5 (1886; Eng. trans., *The Provinces of the Roman Empire,* 2 vol., 1909), a classic; its European sections have now been edited by T.R.S. BROUGHTON, 1968; G.H. STEVENSON, *Roman Provincial Administration till the Age of the Antonines* (1939), sound. For the provincial communities: F.F. ABBOTT and A.C. JOHNSON, *Municipal Administration in the Roman Empire* (1926, reprinted 1968), a standard work; A.H.M. JONES, *The Cities of the Eastern Roman Provinces,* 2nd ed. rev. (1971), essential for the Greek-speaking East. For the armed services: H.M.D. PARKER, *The Roman Legions,* ed. by G.R. WATSON (1958); and C.G. STARR, *The Roman Imperial Navy, 31 B.C.–A.D. 324,* 2nd ed. (1960), both standard works. For general culture: JEAN BEAUJEU, *La Religion romaine à l'apogée de l'Empire* (1955), a very useful survey; J.A. CROOK, *Law and Life of Rome* (1967), sound and readable; W.L. MACDONALD, *The Architecture of the Roman Empire* (1965), excellent; RONALD SYME, *Tacitus,* 2 vol. (1958), a mine of information; J.M.C. TOYNBEE, *The Art of the Romans* (1965), comprehensive and sensible.

The general works treating all aspects of the Later Roman Empire (AD 193 to *c.* 500 in the West, to *c.* 395 in the East) are a little old: *The Cambridge Ancient History,* vol. 12, *The Imperial Crises and Recovery, A.D. 193–324* (1939); and *The Cambridge Medieval History,* vol. 1, *The Christian Roman Empire and the Foundation of the Teutonic Kingdoms* (1936). A detailed (except with regard to events) and excellent work is A.H.M. JONES, *The Later Roman Empire, 284–602: A Social, Economic and Administrative Survey,* 3 vol. (1964). On the problems of civilization and the passage from the Roman to the barbarian worlds, see JOSEPH VOGT, *Der Niedergang Roms* (1965; Eng. trans., *The Decline of Rome: The Metamorphosis of Ancient Civilization,* 1967), a remarkable work in all respects. In the absence of recent works on the Severan dynasty, that of G.J. MURPHY, *The Reign of the Emperor L. Septimius Severus, from the Evidence of the Inscriptions* (1945), assembles the epigraphic sources. L. ROSS TAYLOR, *The Divinity of the Roman Emperor* (1931); and FRANZ CUMONT, *Les Religions orientales dans le paganisme romain,* 4th ed. (1929), are classic and reliable. See also the great literary histories of ALBIN LESKY, *Geschichte der griechischen Literatur,* 2nd ed. (1963; Eng. trans., *A History of Greek Literature,* 1966); and H.J. ROSE, *A Handbook of Latin Literature,* 2nd rev. ed. (1949). The best work on ancient "paideia" is that of H.I. MARROU, *Histoire de l'éducation dans l'antiquité,* 5th ed. (1960; Eng. trans., *A History of Education in Antiquity,* 1956). Economic aspects are discussed in F.M. HEICHELHEIM, *An Ancient Economic History,* vol. 3 (1970); and above all by M.I. ROSTOVTZEFF, *The Social and Economic History of the Roman Empire,* 2nd ed. rev. by P.M. FRASER, 2 vol. (1957). On the transformations of society, C.W. KEYES, *The Rise of the Equities in the Third Century of the Roman Empire* (1915); and the most recent and complete work of JEAN GAGE, *Les Classes sociales dans l'Empire romain* (1964), are recommended.

On municipal life, F.F. ABBOTT and A.C. JOHNSON, *Municipal Administration in the Roman Empire* (1926, reprinted 1968), includes an excellent text and a wide choice of documents. The religious problems created by Constantine's conversion and the birth of the Christian empire are treated in the classic works of ANDRAS ALFOLDI, *The Conversion of Constantine and Pagan Rome* (1948); and of A.H.M. JONES, *Constantine and the Conversion of Europe* (1949). Questions of culture and mind of the 4th century are dealt with in the essays of several scholars and brought together by ARNALDO MOMIGLIANO in *The Conflict Between Paganism and Christianity in the Fourth Century* (1963). A lively and well-documented work is SAMUEL DILL, *Roman Society in the Last Century of the Western Empire,* 2nd ed. (1958); and for the history of ideas, see SANTO MAZZARINO, *La fine del mondo antico* (1959; Eng. trans., *The End of the Ancient World,* 1966). The very important work of VOGT (cited above) gives an excellent bibliography (pp. 319–328) on all the problems treated in this section.

The Classical Greek Dramatists: Aeschylus, Sophocles, Euripides, and Aristophanes

Of the literature of ancient Greece only a small portion survives. Yet much of it remains important, not only because of its high quality but also because Western literature is partly based on forms that were of Greek invention. Two such literary forms are tragedy (*tragōidia;* "goat-song") and comedy (*komōidia;* "revel-song"), both of which presumably originated in ancient Greek rituals of celebration and sacrifice. Both tragedy and comedy were performed at Athens at the spring festival of the god Dionysus (the Great Dionysia). The only extant Greek tragedies are a total of 35 plays written by the Athenian poet-playwrights Aeschylus, Sophocles, and Euripides. In the works of these three men Greek tragedy can be seen both at its creation and at the height of its formal and artistic development. Similarly, the 11 extant complete plays by the Athenian playwright Aristophanes are the only surviving examples of the very first stage in the development of comedy in the Western world, known as Old Comedy. This article treats the lives, works, and achievements of these four classical dramatists. For treatment of the origins, forms, and development of ancient Greek drama and dramatic production, see GREEK LITERATURE; THEATRE, THE HISTORY OF WESTERN; THEATRICAL PRODUCTION.

The article is divided into the following sections:

Aeschylus 342
  Life and career 342
  Dramatic and literary achievements 342
  The plays 343
Sophocles 344
  Life and career 344
  Dramatic and literary achievements 344
  The plays 345
Euripides 346
  Life and career 346
  Dramatic and literary achievements 346
  The plays 347
Aristophanes 348
  Life and career 349
  Dramatic and literary achievements 349
  The plays 349
Major works 350
Bibliography 350

Aeschylus

LIFE AND CAREER

Aeschylus was the first of classical Athens' three great writers of tragedy. He grew up in the turbulent period when the Athenian democracy, having thrown off its tyranny (the absolute rule of one man), had to prove itself against both self-seeking politicians at home and invaders from abroad. Aeschylus himself took part in his city's first struggles against the invading Persians. Later Greek chroniclers believed that Aeschylus was 35 years old in 490 BC when he participated in the Battle of Marathon, in which the Athenians first repelled the Persians; if this is true it would place his birth in 525 BC. Aeschylus' father's name was Euphorion, and the family probably lived at Eleusis (west of Athens).

Aeschylus was a notable participant in Athens' major dramatic competition, the Great Dionysia, which was a part of the festival of Dionysus. Every year at this festival, each of three dramatists would produce three tragedies, which either could be unconnected in plot sequence or could have a connecting theme. This trilogy was followed by a satyr play, which was a kind of lighthearted burlesque. Aeschylus is recorded as having participated in this competition, probably for the first time, in 499 BC. He won his first victory in the theatre in the spring of 484 BC. In the meantime, he had fought and possibly been wounded at Marathon, and Aeschylus singled out his participation in this battle years later for mention on the verse epitaph he wrote for himself. Aeschylus' brother was killed in this battle. In 480 the Persians again invaded Greece, and once again Aeschylus saw service, fighting at the battles of Artemisium and Salamis. His responses to the Persian invasion found expression in his play *Persians,* the earliest of his works to survive. This play was produced in the competition of the spring of 472 BC and won first prize.

Around this time Aeschylus is said to have visited Sicily to present *Persians* again at the tyrant Hieron I's court in Syracuse. Aeschylus' later career is a record of sustained dramatic success, though he is said to have suffered one memorable defeat, at the hands of the novice Sophocles, whose entry at the Dionysian festival of 468 BC was victorious over the older poet's entry. Aeschylus recouped the loss with victory in the next year, 467, with his Oedipus trilogy (of which the third play, *Seven Against Thebes,* survives). After producing his masterpiece, the *Oresteia* trilogy, in 458, Aeschylus went to Sicily again. The chronographers recorded Aeschylus' death at Gela (on Sicily's south coast) in 456/455, aged 69. A ludicrous story that he was killed when an eagle dropped a tortoise on his bald pate was presumably fabricated by a later comic writer. At Gela he was accorded a public funeral, with sacrifices and dramatic performances held at his grave, which subsequently became a place of pilgrimage for writers.

Aeschylus wrote approximately 90 plays, including satyr plays as well as tragedies; of these, about 80 titles are known. Only seven tragedies have survived entire. One account, perhaps based on the official lists, assigns Aeschylus 13 first prizes, or victories; this would mean that well over half of his plays won, since sets of four plays rather than separate ones were judged. According to the philosopher Flavius Philostratus, Aeschylus was known as the "Father of Tragedy." Aeschylus' two sons also achieved prominence as tragedians. One of them, Euphorion, won first prize in his own right in 431 BC over Sophocles and Euripides.

DRAMATIC AND LITERARY ACHIEVEMENTS

Aeschylus' influence on the development of tragedy was fundamental. Previous to him, Greek drama was limited to one actor and a chorus engaged in a largely static recitation. (The chorus was a group of actors who responded to and commented on the main action of a play with song, dance, and recitation.) The actor could assume different roles by changing masks and costumes, but he was limited to engaging in dialogue only with the chorus. By adding a second actor with whom the first could converse, Aeschylus vastly increased the drama's possibilities for dialogue and dramatic tension and allowed more variety and freedom in plot construction. Although the dominance of the chorus in early tragedy is ultimately only hypothesis, it is probably true that, as Aristotle says in his *Poetics,* Aeschylus "reduced the chorus' role and made the plot the leading actor." Aeschylus was an innovator in other ways as well. He made good use of stage settings and stage machinery, and some of his works were noted for their spectacular scenic effects. He also designed costumes, trained his choruses in their songs and dances, and probably acted in most of his own plays, this being the usual practice among Greek dramatists.

Formal innovations

But Aeschylus' formal innovations account for only part of his achievement. His plays are of lasting literary value

in their majestic and compelling lyrical language, in the intricate architecture of their plots, and in the universal themes which they explore so honestly. Aeschylus' language in both dialogue and choral lyric is marked by force, majesty, and emotional intensity. He makes bold use of compound epithets, metaphors, and figurative turns of speech, but this rich language is firmly harnessed to the dramatic action rather than used as mere decoration. It is characteristic of Aeschylus to sustain an image or group of images throughout a play; the ship of state in *Seven Against Thebes,* the birds of prey in *Suppliants,* the snare in *Agamemnon.* More generally, Aeschylus deploys throughout a play or trilogy of plays several leading motifs that are often associated with a particular word or group of words. In the *Oresteia,* for example, such themes as wrath, mastery, persuasion, and the contrasts of light and darkness, of dirge and triumphal song, run throughout the trilogy. This sort of dramatic orchestration as applied to careful plot construction enabled Aeschylus to give Greek drama a more truly artistic and intellectual form.

Aeschylean tragedy deals with the plights, decisions, and fates of individuals with whom the destiny of the community or state is closely bound up; in turn, both individual and community stand in close relation to the gods. Personal, social, and religious issues are thus integrated, as they still were in the Greek civilization of the poet's time. Theodicy (*i.e.,* the justifying of the gods' ways to men) was in some sense the concern of Aeschylus, though it might be truer to say that he aimed through dramatic conflict to throw light on the nature of divine justice. Aeschylus and his Greek contemporaries believed that the gods begrudged human greatness and sent infatuation on a man at the height of his success, thus bringing him to disaster. Man's infatuated act was frequently one of impiety or pride (*hubris*), for which his downfall could be seen as a just punishment. In this scheme of things, divine jealousy and eternal justice formed the common fabric of a moral order of which Zeus, supreme among the gods, was the guardian.

But the unjust are not always punished in their lifetime; it is upon their descendants that justice may fall. It was this tradition of belief in a just Zeus and in hereditary guilt that Aeschylus received, and which is evinced in many of his plays. The simplest illustration of this is in *Persians,* in which Xerxes and his invading Persians are punished for their own offenses. But in a play such as *Agamemnon,* the issues of just punishment and moral responsibility, of human innocence and guilt, of individual freedom versus evil heredity and divine compulsion are more complex and less easily disentangled, thus presenting contradictions which still baffle the human intellect.

Theme of divine justice

Finally, to Aeschylus, divine justice uses human motives to carry out its decrees. Chief among these motives is the desire for vengeance, which was basic to the ancient Greek scheme of values. In the one complete extant trilogy, the *Oresteia,* this notion of vengeance or retaliation is dominant. Retaliation is a motive of Agamemnon, Clytemnestra, Aegisthus, and Orestes. But significantly, the chain of retaliatory murder that pursues Agamemnon and his family ends not by a perfect balance of blood guilt, not by a further perpetuation of violence, but rather through reconciliation and the rule of law as established by Athena and the Athenian courts of justice.

Aeschylus is almost unequaled in writing tragedy that, for all its power of depicting evil and the fear and consequences of evil, ends, as in the *Oresteia,* in joy and reconciliation. Living at a time when the Greek people still truly felt themselves surrounded by the gods, Aeschylus nevertheless had a capacity for detached and general thought, which was typically Greek and which enabled him to treat the fundamental problem of evil with singular honesty and success.

THE PLAYS

Persians. One of a trilogy of unconnected tragedies presented in 472 BC, *Persians* (Greek *Persai*) is unique among surviving tragedies in that it dramatizes recent history rather than events from the distant age of mythical heroes. The play treats the decisive repulse of the Persians from Greece in 480, in particular their defeat at the Battle of Salamis. The play is set in the Persian capital, where a messenger brings news to the Persian queen of the disaster at Salamis. After attributing the defeat of Persia to both Greek independence and bravery and to the gods' punishment of Persian folly for going outside the bounds of Asia, the play ends with the return of the broken and humiliated Persian king, Xerxes.

Seven Against Thebes. This is the third and only surviving play of a connected trilogy, presented in 467 BC, that dealt with the impious transgressions of Laius and the doom subsequently inflicted upon his descendants. The first play seems to have shown how Laius, king of Thebes, had a son despite the prohibition of the oracle of the god Apollo. In the second play it appears that that son, Oedipus, killed his father and laid a curse on his own two sons, Eteocles and Polyneices. In *Seven Against Thebes* (Greek *Hepta epi Thēbais*) Eteocles is shown leading the defense of the city of Thebes against an invading army led by his brother Polyneices and six chieftains from the south of Greece who are bent on placing Polyneices on the Theban throne. Eteocles assigns defenders to each of six of the seven gates of Thebes; but he insists on fighting at the seventh gate, where his opponent will be Polyneices. There the brothers kill each other, and the Theban royal family is thus exterminated, bringing to an end the horrors set in motion by Laius' defiance of the gods.

Suppliants. This is the first and only surviving play of a trilogy probably put on in 463. It was long believed by scholars that *Suppliants* (Greek *Hiketides;* Latin *Supplices*) was one of Aeschylus' earliest plays because of its archaic structure; its chorus, representing the daughters of Danaus (the Danaïds), takes the leading role in the action. But there is now evidence that the trilogy of which *Suppliants* formed a part was produced in competition with Sophocles, who is first known to have competed in 468. *Suppliants* thus dates presumably from the middle of Aeschylus' career, not from the beginning.

Born in Egypt, though of Greek descent, the Danaïds have fled with their father to Argos in Greece in order to avoid forced marriage with their cousins, the sons of Aegyptus. Pelasgus, the king of Argos, is torn between charity to the Danaïds and anxiety to appease Aegyptus but nobly agrees in the end to grant them asylum. The trilogy as a whole seems to have favourably stressed the saving power of domestic love as contrasted with both the willful virginity of the Danaïds and the unfeeling, violent lust of their cousins.

Oresteia. The *Oresteia* trilogy consists of three closely connected plays, all extant, that were presented in 458 BC. In *Agamemnon* the great Greek king of that name returns triumphant from the siege of Troy, along with his concubine, the Trojan prophetess Cassandra, only to be humiliated and murdered by his fiercely vengeful wife, Clytemnestra. She is driven to this act partly by a desire to avenge the death of her daughter Iphigenia, whom Agamemnon has sacrificed for the sake of the war, partly by her adulterous love for Aegisthus, and partly as agent for the curse brought on Agamemnon's family by the crimes of his father, Atreus. At the play's end Clytemnestra and her lover have taken over the palace and now rule Argos. Many regard this play as one of the greatest Greek tragedies. From its extraordinarily sustained dramatic and poetic power one might single out the fascinating, deceitful richness of Clytemnestra's words and the huge choral songs, which raise in metaphorical and often enigmatic terms the complex of major themes—of theology, politics, and blood relationships—which are elaborated throughout the trilogy.

Oresteia trilogy

Libation Bearers (Greek *Choēphoroi*) is the second play in the trilogy and takes its title from the chorus of women servants who come to pour propitiatory offerings at the tomb of the murdered Agamemnon. At the start of this play Orestes, the son of Agamemnon and Clytemnestra, who was sent abroad as a child, returns as a man to take vengeance upon his mother and her lover for their murder of his father. He is reunited with his sister Electra, and together they invoke the aid of the dead Agamemnon in their plans. Orestes then slays Aegisthus, but Orestes'

subsequent murder of Clytemnestra is committed reluctantly, at the god Apollo's bidding. Orestes' attempts at self-justification then falter and he flees, guilt-wracked, maddened, and pursued by the female incarnations of his mother's curse, the Erinyes (Furies). At this point the chain of vengeance seems interminable.

Eumenides, the title of the third play, means "The Kind Goddesses." The play opens at the shrine of Apollo at Delphi, where Orestes has taken sanctuary from the Furies. At the command of the Delphic oracle, Orestes journeys to Athens to stand trial for his matricide. There the goddess Athena organizes a trial with a jury of citizens. The Furies are his accusers, while Apollo defends him. The jury divides evenly in its vote and Athena casts the tie-breaking vote for Orestes' acquittal. The Furies then turn their vengeful resentment against the city itself, but Athena persuades them, in return for a home and cult, to bless Athens instead and reside there as the "Kind Goddesses" of the play's title. The trilogy thus ends with the cycle of retributive bloodshed ended and supplanted by the rule of law and the justice of the state.

Prometheus Bound. The date of this play (and even its authorship) is disputed, but many scholars regard it as a work of Aeschylus' last years. In *Prometheus Bound* (Greek *Promētheus desmōtēs*) the god Prometheus, who in defiance of Zeus has saved mankind and given them fire, is chained to a remote crag as a punishment ordered by the king of the gods. Despite his isolation Prometheus is visited by the ancient god Oceanus, by a chorus of Oceanus' daughters, by the "cow-headed" Io (another victim of Zeus), and finally by the god Hermes, who vainly demands from Prometheus his knowledge of a secret that could threaten Zeus's power. After refusing to reveal his secret, Prometheus is cast into the underworld for further torture. The drama of the play lies in the clash between the irresistible power of Zeus and the immovable will of Prometheus, who has been rendered still more stubborn by Io's misfortunes at the hands of Zeus. The most striking and controversial aspect of the play is its depiction of Zeus as a tyrant. Prometheus himself has proved to be for later ages an archetypal figure of defiance against tyrannical power, a role exemplified in Percy Bysshe Shelley's poem *Prometheus Unbound* (1820).

(A.J.P./O.T.)

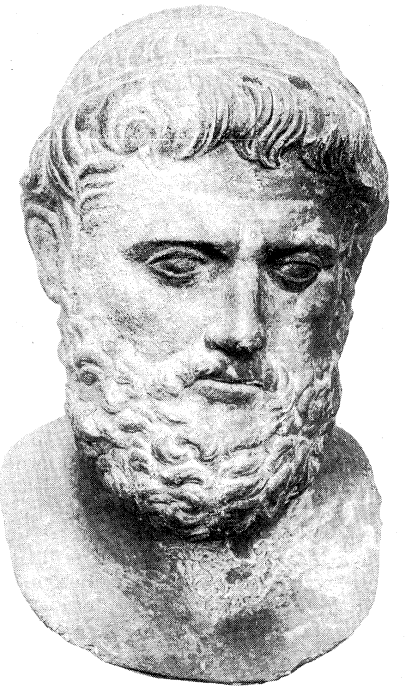
Sophocles

LIFE AND CAREER

The second of classical Athens' three great writers of tragedy, Sophocles was the younger contemporary of Aeschylus and the older contemporary of Euripides. He was born about 496 BC at Colonus, a village outside the walls of Athens. His father, Sophillus, was a wealthy manufacturer of armour, and Sophocles himself received a good education. Because of his beauty of physique, his athletic prowess, and his skill in music, he was chosen in 480, when he was 16, to lead the paean (choral chant to a god) celebrating the decisive Greek sea victory over the Persians at the Battle of Salamis. The relatively meagre information about Sophocles' civic life suggests that he was a popular favourite who participated actively in his community and exercised outstanding artistic talents. In 442 he served as one of the treasurers responsible for receiving and managing tribute money from Athens' subject-allies in the Delian League. In 440 he was elected one of the 10 *stratēgoi* (high executive officials who commanded the armed forces) as a junior colleague of Pericles. Sophocles later served as *stratēgos* perhaps twice again. In 413, then aged about 83, Sophocles was a *proboulos,* one of 10 advisory commissioners who were granted special powers and were entrusted with organizing Athens' financial and domestic recovery after its terrible defeat at Syracuse in Sicily. Sophocles' last recorded act was to lead a chorus in public mourning for his deceased rival, Euripides, before the festival of 406. He died that same year.

Public service

These few facts are about all that is known of Sophocles' life. They imply steady and distinguished attachment to Athens, its government, religion, and social forms. Sophocles was wealthy from birth, highly educated, noted for his grace and charm, on easy terms with the leading families, a personal friend of prominent statesmen, and in many ways fortunate to have died before the final surrender of Athens to Sparta in 404. In one of his last plays, *Oedipus at Colonus,* he still affectionately praises both his own birthplace and the great city itself.



Sophocles, bronze bust copied from a Greek original, 340–330 BC. In the Museo Archeologico, Florence.
Alinari—Art Resource/EB Inc.

Sophocles won his first victory at the Dionysian dramatic festival in 468, however, defeating the great Aeschylus in the process. This began a career of unparalleled success and longevity. In total, Sophocles wrote 123 dramas for the festivals. Since each author who was chosen to enter the competition usually presented four plays, this means he must have competed about 30 times. Sophocles won perhaps as many as 24 victories, compared to 13 for Aeschylus and four for Euripides, and indeed he may have never received lower than second place in the competitions he entered.

DRAMATIC AND LITERARY ACHIEVEMENTS

Ancient authorities credit Sophocles with several major and minor dramatic innovations. Among the latter is his invention of some type of "scene paintings" or other pictorial prop to establish locale or atmosphere. He also may have increased the size of the chorus from 12 to 15 members. Sophocles' major innovation was his introduction of a third actor into the dramatic performance. It had previously been permissible for two actors to "double" (*i.e.,* assume other roles during a play), but the addition of a third actor onstage enabled the dramatist both to increase the number of his characters and widen the variety of their interactions. The scope of the dramatic conflict was thereby extended, plots could be more fluid, and situations could be more complex.

The typical Sophoclean drama presents a few characters, impressive in their determination and power and possessing a few strongly drawn qualities or faults that combine with a particular set of circumstances to lead them inevitably to a tragic fate. Sophocles develops his characters' rush to tragedy with great economy, concentration, and dramatic effectiveness, creating a coherent, suspenseful situation whose sustained and inexorable onrush came to epitomize the tragic form to the classical world. Sophocles emphasizes that most people lack wisdom, and he presents truth in collision with ignorance, delusion, and folly. Many scenes dramatize flaws or failure in thinking (deceptive reports and rumours, false optimism, hasty judgment,

Typical dramatic framework

madness). The chief character does something involving grave error; this affects others, each of whom reacts in his own way, thereby causing the chief agent to take another step toward ruin—his own and that of others as well. Equally important, those who are to suffer from the tragic error usually are present at the time or belong to the same generation. It was this more complex type of tragedy that demanded a third actor. Sophocles thus abandoned the spacious Aeschylean framework of the connected trilogy and instead comprised the entire action in a single play. From his time onward, "trilogy" usually meant no more than three separate tragedies written by the same author and presented at the same festival.

Sophocles' language responds flexibly to the dramatic needs of the moment; it can be ponderously weighty or swift-moving, emotionally intense or easygoing, highly decorative or perfectly plain and simple. His mastery of form and diction was highly respected by his contemporaries. Sophocles has also been universally admired for the sympathy and vividness with which he delineates his characters; especially notable are his tragic women, such as Electra and Antigone. Few dramatists have been able to handle situation and plot with more power and certainty; the frequent references in the *Poetics* to Sophocles' *Oedipus the King* show that Aristotle regarded this play as a masterpiece of construction, and few later critics have dissented. Sophocles is also unsurpassed in his moments of high dramatic tension and in his revealing use of tragic irony.

The criticism has been made that Sophocles was a superb artist and nothing more; he grappled neither with religious problems as Aeschylus had nor with intellectual ones as Euripides had done. He accepted the gods of Greek religion in a spirit of unreflecting orthodoxy, and he contented himself with presenting human characters and human conflicts. But it should be stressed that to Sophocles "the gods" appear to have represented the natural forces of the universe to which human beings are unwittingly or unwillingly subject. To Sophocles, human beings live for the most part in dark ignorance because they are cut off from these permanent, unchanging forces and structures of reality. Yet it is pain, suffering, and the endurance of tragic crisis that can bring people into valid contact with the universal order of things. In the process, a person can become more genuinely human, more genuinely himself.

Conception of life

THE PLAYS

Only seven of Sophocles' tragedies survive in their entirety, along with 400 lines of a satyr play, numerous fragments of plays now lost, and 90 titles. All seven of the complete plays are works of Sophocles' maturity, but only two of them, *Philoctetes* and *Oedipus at Colonus,* have fairly certain dates. *Ajax* is generally regarded as the earliest of the extant plays. Some evidence suggests that *Antigone* was first performed in 442 or 441 BC. *Philoctetes* was first performed in 409, when Sophocles was 90 years old, and *Oedipus at Colonus* was said to have been produced after Sophocles' death by his grandson.

Ajax. The entire plot of *Ajax* (Greek *Aias mastigophoros*) is constructed around Ajax, the mighty hero of the Trojan War whose pride drives him to treachery and finally to his own ruin and suicide some two-thirds of the way through the play. Ajax is deeply offended at the award of the prize of valour (the dead Achilles' armour) not to himself but to Odysseus. Ajax thereupon attempts to assassinate Odysseus and the contest's judges, the Greek commanders Agamemnon and Menelaus, but is frustrated by the intervention of the goddess Athena. He cannot bear his humiliation and throws himself on his own sword. Agamemnon and Menelaus order that Ajax' corpse be left unburied as punishment. But the wise Odysseus persuades the commanders to relent and grant Ajax an honourable burial. In the end Odysseus is the only person who seems truly aware of the changeability of human fortune.

Antigone. Antigone is the daughter of Oedipus, the former king of Thebes. She is willing to face the capital punishment that has been decreed by her uncle Creon, the new king, as the penalty for anyone burying her brother Polyneices. (Polyneices has just been killed attacking Thebes, and it is as posthumous punishment for this attack that Creon has forbidden the burial of his corpse.) Obeying all her instincts of love, loyalty, and humanity, Antigone defies Creon and dutifully buries her brother's corpse. Creon, from conviction that reasons of state outweigh family ties, refuses to commute Antigone's death sentence. By the time Creon is finally persuaded by the prophet Tiresias to relent and free Antigone, she has killed herself in her prison cell. Creon's son, Haemon, kills himself out of love and sympathy for the dead Antigone, and Creon's wife, Eurydice, then kills herself out of grief over these tragic events. At the play's end Creon is left desolate and broken in spirit. In his narrow and unduly rigid adherence to his civic duties, Creon has defied the gods through his denial of humanity's common obligations toward the dead. The play thus concerns the conflicting obligations of civic versus personal loyalties and religious mores.

Trachinian Women. This play centres on the efforts of Deianeira to win back the wandering affections of her husband, Heracles, who is away on one of his heroic missions and who has sent back his latest concubine, Iole, to live with his wife at their home in Trachis. The love charm Deianeira uses on Heracles turns out to be poisonous, and she kills herself upon learning of the agony she has caused her husband. Thus, in *Trachinian Women* (Greek *Trachiniai*) Heracles' insensitivity (in sending his mistress to share his wife's home) and Deianeira's ignorance result in domestic tragedy.

Oedipus the King. The plot of *Oedipus the King* (Greek *Oidipous Tyrannos;* Latin *Oedipus Rex*) is a structural marvel that marks the summit of classical Greek drama's formal achievements. The play's main character, Oedipus, is the wise, happy, and beloved ruler of Thebes. Though hot-tempered, impatient, and arrogant at times of crisis, he otherwise seems to enjoy every good fortune. But Oedipus mistakenly believes that he is the son of King Polybus of Corinth and his queen. He became the ruler of Thebes because he rescued the city from the Sphinx by answering its riddle correctly, and so was awarded the city's widowed queen, Jocasta. Before overcoming the Sphinx, Oedipus left Corinth forever because the Delphic oracle had prophesied to him that he would kill his father and marry his mother. While journeying to Thebes from Corinth, Oedipus encountered at a crossroads an old man accompanied by five servants. Oedipus got into an argument with him and in a fit of arrogance and bad temper killed the old man and four of his servants.

The play opens with the city of Thebes stricken by a plague and its citizens begging Oedipus to find a remedy. He consults the Delphic oracle, which declares that the plague will cease only when the murderer of Jocasta's first husband, King Laius, has been found and punished for his deed. Oedipus resolves to find Laius' killer, and much of the rest of the play centres upon the investigation he conducts in this regard. In a series of tense, gripping, and ominous scenes Oedipus' investigation turns into an obsessive reconstruction of his own hidden past as he begins to suspect that the old man he killed at the crossroads was none other than Laius. Finally, Oedipus learns that he himself was abandoned to die as a baby by Laius and Jocasta because they feared a prophecy that their infant son would kill his father; that he survived and was adopted by the ruler of Corinth, but in his maturity he has unwittingly fulfilled the Delphic oracle's prophecy of him; that he has indeed killed his true father, married his own mother, and begot children who are also his own siblings.

Plot of *Oedipus the King*

Jocasta hangs herself when she sees this shameful web of incest, parricide, and attempted child murder, and the guilt-stricken Oedipus then sticks needles into his eyes, blinding himself. Sightless and alone, he is now blind to the world around him but finally cognizant of the terrible truth of his own life.

Electra. As in Aeschylus' *Libation Bearers,* the action in *Electra* (Greek *Ēlektra*) follows the return of Orestes to kill his mother, Clytemnestra, and her lover Aegisthus in retribution for their murder of Orestes' father, Agamemnon. In this play, however, the main focus is on Orestes' sister Electra and her anguished participation in her brother's plans. To gain admittance to the palace and thus be

Portrait of Electra

able to execute his revenge, Orestes spreads false news of his own death. Believing this report, the despairing Electra unsuccessfully tries to enlist her sister Chrysothemis in an attempt to murder their mother. In a dramatic scene, Orestes then enters in disguise and hands Electra the urn that is supposed to contain his own ashes. Moved by his sister's display of grief, Orestes reveals his true identity to her and then strikes down his mother and her lover. Electra's triumph is thus complete. In the play Electra is seen passing through the whole range of human emotions—from passionate love to cruel hatred, from numb despair to wild joy. There is debate over whether the play depicts virtue triumphant or, rather, portrays a young woman incurably twisted by years of hatred and resentment.

Philoctetes. In *Philoctetes* (Greek *Philoktētēs*) the Greeks on their way to Troy have cast away the play's main character, Philoctetes, on the desert island of Lemnos because he has a loathsome and incurable ulcer on his foot. But the Greeks have discovered that they cannot win victory over Troy without Philoctetes and his wonderful bow, which formerly belonged to Heracles. The crafty Odysseus is given the task of fetching Philoctetes by any means possible. Odysseus knows that the resentful Philoctetes will kill him if he can, so he uses the young and impressionable soldier Neoptolemus, son of the dead Achilles, as his agent. Neoptolemus is thus caught between the devious manipulations of Odysseus and the unsuspecting integrity of Philoctetes, who is ready to do anything rather than help the Greeks who abandoned him. For much of the play Neoptolemus sticks to Odysseus' policy of deceit, despite his better nature, but eventually he renounces duplicity to join in friendship with Philoctetes. A supernatural appearance by Heracles then convinces Philoctetes to go to Troy to both win victory and be healed of his disease.

Oedipus at Colonus. In *Oedipus at Colonus* (Greek *Oidipous epi Kolōnō*) the old, blind Oedipus has spent many years wandering in exile after being rejected by his sons and the city of Thebes. Oedipus has been cared for only by his daughters Antigone and Ismene. He arrives at a sacred grove at Colonus, a village close by Athens (and the home of Sophocles himself). There Oedipus is guaranteed protection by Theseus, the noble king of Athens. Theseus does indeed protect Oedipus from the importunate pleadings of his brother-in-law, Creon, for Oedipus to protect Thebes. Oedipus himself rejects the entreaties of his son Polyneices, who is bent on attacking Thebes and whom Oedipus solemnly curses. Finally Oedipus departs to a mysterious death; he is apparently swallowed into the earth of Colonus, where he will become a benevolent power and a mysterious source of defense to the land that has given him final refuge. The play is remarkable for the melancholy, beauty, and power of its lyric odes and for the spiritual and moral authority with which it invests the figure of Oedipus.

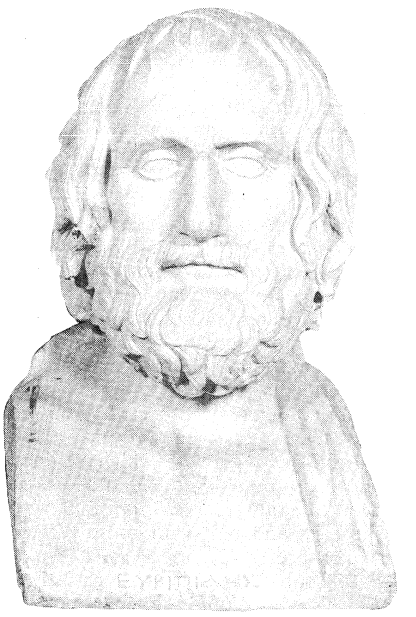
Trackers. Four hundred lines of this satyr play survive. The plot of *Trackers* (Greek *Ichneutai*) is based on two stories about the miraculous early deeds of the god Hermes: that the infant, growing to maturity in a few days, stole cattle from Apollo, baffling discovery by reversing the animals' hoof marks, and that he invented the lyre by fitting strings to a tortoise shell. In this play the trackers are the chorus of satyrs, who are looking for the cattle; they are amusingly dumbfounded at the sound of the new instrument Hermes has invented. Enough of the play survives to give an impression of its style; it is a genial, uncomplicated travesty of the tragic manner, and the antics of the chorus were apparently the chief source of amusement. (T.W./O.T.)

Satyr play

Euripides

LIFE AND CAREER

It is possible to reconstruct only the sketchiest biography of Euripides, the youngest of classical Athens' three great tragedians. Euripides' mother's name was Cleito; his father's name was Mnesarchus or Mnesarchides. One tradition states that his mother was a greengrocer who sold herbs in the marketplace. Aristophanes joked about this in comedy after comedy; but there is better indirect evidence that Euripides came of a well-off family. He was probably born in 484. Euripides first received the honour of being chosen to compete in the dramatic festival in 455, and he won his first victory in 441. Euripides left Athens for good in 408, accepting an invitation from Archelaus, king of Macedonia. He died in Macedonia in 406.



Euripides, marble herm copied from a Greek original, *c.* 340–330 BC. In the Museo Archeologico Nazionale, Naples.

By courtesy of the Soprintendenza alle Antichita della Campania, Naples

Euripides' only known public activity was his service on a diplomatic mission to Syracuse in Sicily. He was passionately interested in ideas, however, and owned a large library. He is said to have associated with Protagoras, Anaxagoras, and other Sophists and philosopher-scientists. His acquaintance with new ideas brought him restlessness rather than conviction, however, and his questioning attitude toward traditional Greek religion is reflected in some of his plays. Of Euripides' private life, little can be said. Later tradition invented for him a spectacularly disastrous married life. It is known that he had a wife called Melito and produced three sons. One of these was something of a poet and produced the *Bacchants* after his father's death. He may also have completed his father's unfinished play *Iphigenia at Aulis.*

Uncertain biographical accounts

The ancients knew of 92 plays composed by Euripides. Nineteen plays are extant, if one of disputed authorship is included. At only four festivals was Euripides awarded the first prize—the fourth posthumously, for the tetralogy that included *Bacchants* and *Iphigenia at Aulis.* As Sophocles won perhaps as many as 24 victories, it is clear that Euripides was comparatively unsuccessful. More to the point is that on more than 20 occasions Euripides was chosen, out of all contestants, to be one of the three laureates of the year. Furthermore, the regularity with which Aristophanes parodied him is proof enough that Euripides' work commanded attention. It is often said that disappointment at his plays' reception in Athens was one of the reasons for his leaving his native city in his old age; but there are other reasons why an old poet might have left Athens in the 23rd year of the Peloponnesian War.

DRAMATIC AND LITERARY ACHIEVEMENTS

Euripides' plays exhibit his iconoclastic, rationalizing attitude toward both religious belief and the ancient legends and myths that formed the traditional subject matter for Greek drama. These legends seem to have been for him a mere collection of stories without any particular authority. He also apparently rejected the gods of Homeric theology, whom he frequently depicts as irrational, petulant, and singularly uninterested in meting out "divine justice."

That the gods are so often presented on the stage by Euripides is partly due to their convenience as a source of information that could not otherwise be made available to the audience.

Given this attitude of sophisticated doubt on his part, Euripides invents protagonists who are quite different from the larger-than-life characters drawn with such conviction by Aeschylus and Sophocles. They are, for the most part, commonplace, down-to-earth men and women who have all the flaws and vulnerabilities ordinarily associated with human beings. Furthermore, Euripides makes his characters express the doubts, the problems and controversies, and in general the ideas and feelings of his own time. They sometimes even take time off from the dramatic action to debate each other on matters of current philosophical or social interest.

Euripides differed from Aeschylus and Sophocles in making his characters' tragic fates stem almost entirely from their own flawed natures and uncontrolled passions. Chance, disorder, and human irrationality and immorality frequently result not in an eventual reconciliation or moral resolution but in apparently meaningless suffering that is looked upon with indifference by the gods. The power of this type of drama lies in the frightening and ghastly situations it creates and in the melodramatic, even sensational, emotional effects of its characters' tragic crises.

Psychological realism

Given this strong strain of psychological realism, Euripides shows moments of brilliant insight into his characters, especially in scenes of love and madness. His depictions of women deserve particular attention; it is easy to extract from his plays a long list of heroines who are fierce, treacherous, or adulterous, or all three at once. Misogyny is altogether too simple an explanation here, although Euripides' reputation in his own day was that of a woman hater, and a play by Aristophanes, *Women at the Thesmophoria,* comically depicts the indignation of the Athenian women at their portrayal by Euripides.

The chief structural peculiarities of Euripides' plays are his use of prologues and of the providential appearance of a god (deus ex machina) at the play's end. Almost all of the plays start with a monologue that is in effect a bare chronicle explaining the situation and characters with which the action begins. Similarly, the god's epilogue at the end of the play serves to reveal the future fortunes of the characters. This latter device has been criticized as clumsy or artificial by modern authorities, but it was presumably more palatable to the audiences of Euripides' own time. Another striking feature of his plays is that over time Euripides found less and less use for the chorus; in his successive works it tends to grow detached from the dramatic action.

The word habitually used in antiquity to describe Euripides' ordinary style of dramatic speech is *lalia* ("chatter"), alluding probably both to its comparatively light weight and to the volubility of his characters of all classes. Notwithstanding this, Euripides' lyrics at times have considerable charm and sweetness. In the works written after 415 BC his lyrics underwent a change, becoming more emotional and luxuriant. At its worst this style is hardly distinguishable from Aristophanes' parody of it in his comedy *Frogs,* but where frenzied emotion is appropriate, as in the tragedy *Bacchants,* Euripides' songs are unsurpassed in their power and beauty.

The tragicomedies

During the last decade of his career Euripides began to write "tragedies" that might actually be called romantic dramas, or tragicomedies with happy endings. These plays have a highly organized structure leading to a recognition scene in which the discovery of a character's true identity produces a complete change in the situation, and in general a happy one. Extant plays in this style include *Ion, Iphigenia Among the Taurians,* and *Helen.* Plays of the tragicomedy type seem to anticipate the New Comedy of the 4th century BC.

The fame and popularity of Euripides eclipsed that of Aeschylus and Sophocles in the cosmopolitan Hellenistic period. The austere, lofty, essentially political and "religious" tragedy of Aeschylus and Sophocles had less appeal than that of Euripides, with its more accessible realism and its obviously emotional, even sensational, effects. Euripides thus became the most popular of the three for revivals of his plays in later antiquity; this is probably why at least 18 of his plays have survived compared to seven each for Aeschylus and Sophocles, and why the extant fragmentary quotations from his works are more numerous than those of Aeschylus and Sophocles put together.

THE PLAYS

The dates of production of nine of Euripides' plays are known with some certainty from evidence that goes back to the official Athenian records. Those plays whose dates are prefixed by *c.* can be dated to within a few years by the internal evidence of Euripides' changing metrical techniques.

Alcestis. Though tragic in form, *Alcestis* (438 BC; Greek *Alkēstis*) ends happily and took the place of the satyr play that normally followed the three tragedies. King Admetus is doomed to die shortly, but he will be allowed a second life if he can find someone willing to die in his place. His wife, Alcestis, voluntarily dies in place of her husband, who sees too late that the fact and manner of her dying will blight his life. But Admetus' old friend Heracles shows up and rescues Alcestis from the clutches of Death, restoring her to her happy and relieved husband.

Medea. One of Euripides' most powerful and best known plays, *Medea* (431 BC; Greek *Mēdeia*) is a remarkable study of the mistreatment of a woman and of her ruthless revenge. The Colchian princess Medea has been taken by the hero Jason to be his wife. They have lived happily for some years at Corinth and have two sons. But then Jason casts Medea off and decides to marry the Princess of Corinth. Medea is determined on revenge, and after a dreadful mental struggle between her passionate sense of injury and her love for her children, she decides to punish her husband by murdering both the Corinthian princess and their own sons, thereby leaving her husband to grow old with neither wife nor child. She steels herself to commit these deeds and then escapes in the chariot of her grandfather, the sun-god Helios, leaving Jason without even the satisfaction of punishing her for her crimes. Euripides succeeds in evoking sympathy for the figure of Medea, who becomes to some extent a representative of women's oppression in general.

Children of Heracles. The plot of *Children of Heracles* (430 BC; Greek *Hērakleidai*) concerns the Athenians' defense of the young children of the dead Heracles from the murderous intentions of King Eurystheus of Argos. The play is basically a simple glorification of Athens.

Hippolytus. In *Hippolytus* (428 BC; Greek *Hippolytos*) Aphrodite, the goddess of love and sexual desire, destroys Hippolytus, a lover of outdoor sports who is repelled by sexual passion and who is instead devoted to the virgin huntress Artemis. Aphrodite makes Phaedra, wife of Theseus, the king of Athens, fall violently in love with her stepson Hippolytus. Phaedra is deeply ashamed of her illicit passion, but when Hippolytus angrily rejects her love she is so mortified by his denunciation that she cannot forbear from falsely accusing him of rape before she kills herself. Her accusation provokes Theseus into pronouncing a curse on his son that eventually leads to Hippolytus' death. But Artemis reveals Hippolytus' innocence before he dies, and the young man is able to forgive his father, thus freeing Theseus from the dreadful stain of bloodguilt. Given the nature of its plot, the play is remarkable for its propriety.

Andromache. This play is set in the aftermath of the Trojan War. After an exciting beginning marked by strong anti-Spartan feeling, most of the original characters in *Andromache* (*c.* 426 BC) disappear and the interest is dissipated.

Miseries of war

Hecuba. Also set in the aftermath of the Trojan War, *Hecuba* (*c.* 425 BC; Greek *Hekabē*) shows the double disaster that reduces the aged Trojan queen Hecuba, now a widowed slave, by sheer weight of hatred and misery to a mere animal ferocity. Hecuba first loses her daughter Polyxena, who is taken off to be sacrificed to the ghost of Achilles. Hecuba then discovers the corpse of her last son, Polydorus, who has been murdered by his Thracian host, Polymestor. Hecuba eventually persuades the Greek

commander Agamemnon to allow her to take vengeance; she and her women then blind Polymestor and murder his two young sons. Such is the power of misery to deprave, and the play's closing prophecy of Hecuba's future transformation into a bitch seems appropriate.

Suppliants. The title figures of *Suppliants* (*c.* 423 BC; Greek *Hiketides;* Latin *Supplices*) are the mothers of the Argive leaders who have been killed while attacking Thebes. The bodies of their sons have been left unburied by the Thebans, and they eventually persuade the Athenians to recover them. It is disputed whether the play is a straightforward eulogy of Athens and its democracy, or whether its sentiments are being expressed ironically.

Electra. The title character of *Electra* (*c.* 418 BC; Greek *Ēlektra*) and her brother Orestes murder their mother, Clytemnestra, in retribution for her murder of their father, Agamemnon. Electra herself is portrayed as a frustrated and resentful woman who finally lures her mother to her death by appealing to her maternal instincts. After the horrible murder both Electra and her reluctant accomplice Orestes are consumed by remorse. This is a bitterly realistic and antiheroic play that draws a disturbingly convincing portrait of both Electra's sufferings and her unattractive personality.

Madness of Heracles. The title character of *Madness of Heracles* (*c.* 416 BC; Greek *Hēraklēs mainomenos;* Latin *Hercules furens*) is temporarily driven mad by the goddess Hera and kills his wife and children. Subsequently Heracles recovers his reason and, after recovering from suicidal despair, is taken to spend an honourable retirement at Athens.

Trojan Women. The setting of *Trojan Women* (415 BC; Greek *Trōades*) is the time immediately after the taking of Troy, and the play treats the sufferings of the wives and children of the city's defeated leaders, in particular the old Trojan queen Hecuba and her children. Hecuba's daughter Cassandra is taken off to be the concubine of Agamemnon, and then her daughter-in-law Andromache is led off to be the slave of Neoptolemus. Andromache's son Astyanax is taken from her to be hurled to his death from the walls of Troy. Finally, as Troy goes up in flames, Hecuba and the other Trojan women are taken off to the ships to face slavery in Greece. This play is a famous and powerful indictment of the barbarous cruelties of war. It was first produced only months after the Athenians captured the city-state of Melos, butchering its men and reducing its women to slavery, and the *Trojan Women*'s mood may well have been influenced by the Athenians' atrocities and the Melians' fate, which are both mirrored in the play.

Plays with recognition scenes

Ion. This tragicomedy's sombre action is reversed by a recognition scene. In *Ion* (*c.* 413 BC), Creusa, the queen of Athens, is married to an immigrant king, Xuthus, but the couple do not have any children. Years before, the Queen was raped by the god Apollo but abandoned the subsequent child. The boy Ion has grown up as a temple slave at Delphi, where the play is set. When they meet, mother and son feel a strong affinity, but when the Delphic oracle says the boy is the son of Xuthus, the Queen in her despairing childlessness plots to kill the young stranger who threatens to take over her inheritance. At the last minute they recognize each other by means of the cradle Creusa had long ago left with her baby. The play has a superficially satisfactory ending, but its portrayal of human suffering and of divine carelessness and mendacity is tinged with darker feelings.

Iphigenia Among the Taurians. This is another tragicomedy, composed chiefly of a recognition scene followed by a clever escape. The title character of *Iphigenia Among the Taurians* (*c.* 413 BC; Greek *Iphigeneia en Taurois;* Latin *Iphigenia in Tauris*) has been saved by the goddess Artemis from sacrifice by her father and now serves the goddess' temple at Tauris in Thrace. Iphigenia's brother Orestes is captured by the local tyrant and is delivered to her for sacrifice. She recognizes him, however, and after some exciting mishaps they manage to escape from Tauris with the help of Athena.

Helen. In this frankly light play, Euripides deflates one of the best known "facts" of Greek mythology, that Helen ran off adulterously with Paris to Troy. In *Helen* (412 BC; Greek *Helenē*) only a phantom went with Paris to Troy, and the real Helen pines faithfully in Egypt. When Menelaus on his way home from Troy is shipwrecked in Egypt, he is baffled by the duplicate Helen until the evaporation of the phantom allows his reunion with the real one. The pair then escape from the King of Egypt, who is keen to marry Helen, by an amusing artifice.

Phoenician Women. This is a diverse, many-characteried play whose original version has been tampered with. *Phoenician Women* (*c.* 409 BC; Greek *Phoinissai*) is set at Thebes and concerns the mutual slaughter of the two sons of Oedipus, Eteocles and Polyneices.

Orestes. In this play Euripides makes nonsense of the old story of Orestes' murder of his mother, Clytemnestra, by setting the play in a world where courts of law already exist. In *Orestes* (408 BC), the main character, his sister Electra, and his cousin and friend Pylades are condemned to death by the men of Argos for the murder. Their uncle Menelaus is too spineless to defend them, and they are finally reduced to plotting to kill Menelaus' wife, Helen, and abduct her innocent daughter. This chaos of violence and attempted murder is only resolved by the deus ex machina Apollo, who appears and restores harmony at the end of the play.

Iphigenia at Aulis. The Greek fleet is becalmed at Aulis and is thus unable to convey the expeditionary force against Troy. Agamemnon learns that he must sacrifice his daughter Iphigenia as a means of appeasing the goddess Artemis, who has caused the unfavourable weather. Agamemnon lures his daughter into coming to Aulis to be sacrificed by pretending that she will marry Achilles. Once the truth is out, Iphigenia, after begging pathetically for her life, goes willingly to her death. Though incomplete and corrupted by later adapters, *Iphigenia at Aulis* (*c.* 406 BC; Greek *Iphigeneia en Aulidi*) is a fine tragedy whose realistic atmosphere is heightened by several subtle and poignant scenes between its main characters.

Bacchants. This play is regarded by many as Euripides' masterpiece. In *Bacchants* (*c.* 406 BC; Greek *Bakchai;* Latin *Bacchae*) the god Dionysus arrives in Greece from Asia intending to introduce his orgiastic worship there. He is disguised as a charismatic young Asian holy man and is accompanied by his women votaries, who make up the play's chorus. He expects to be accepted first in Thebes, but the Thebans reject his divinity and refuse to worship him, and the city's young king, Pentheus, tries to arrest him. In the end Dionysus drives Pentheus insane and leads him to the mountains, where Pentheus' own mother, Agave, and the women of Thebes in a bacchic frenzy tear him to pieces. Agave returns to Thebes triumphant carrying Pentheus' head, and her father, Cadmus, has to lead her back to sanity and recognition. The play shows how the liberating and ecstatic side of the Dionysiac religion must be balanced against the dangerous irresponsibility that goes with the Dionysiac loss of reason and self-consciousness.

Late masterpiece

Cyclops. *Cyclops* (Greek *Kyklōps*) is the only complete surviving satyr play. The play's cowardly, lazy satyrs with their disgraceful old father Silenus are slaves of the man-eating one-eyed Cyclops Polyphemus in Sicily. Odysseus arrives, driven to Sicily by adverse weather, and eventually succeeds (as in Homer's *Odyssey*) in blinding the Cyclops. He thus enables the Cyclops' victims to escape.

(H.D.F.K./O.T.)

Aristophanes

The most famous of all comic dramatists of ancient Greece, Aristophanes is also the one whose works have been preserved in the greatest quantity. He is the only extant representative of the Old Comedy, that is, of the phase of comic dramaturgy in which chorus, mime, and burlesque still played a considerable part and which was characterized by bold fantasy, merciless invective and outrageous satire, unabashedly licentious humour, and a marked freedom of political criticism. But Aristophanes belongs to the end of this phase, and, indeed, his last extant play, which has no choric element at all, may well

be regarded as the only extant specimen of the short-lived Middle Comedy, which, before the end of the 4th century BC, was to be superseded in turn by the milder and more realistic social satire of the New Comedy.

LIFE AND CAREER

Little is known about the life of Aristophanes, and most of the known facts are derived from references in his own plays. Born *c.* 450 BC, he was an Athenian citizen belonging to the *deme,* or clan, named Pandionis, but his actual birthplace is uncertain. (The fact that he or his father, Philippus, owned property on the island of Aegina may have been the cause of an accusation by his fellow citizens that he was not of Athenian birth.) He began his dramatic career in 427 BC with a play, the *Daitaleis* (*The Banqueters*), which appears, from surviving fragments, to have been a satire on his contemporaries' educational and moral theories. He is thought to have written about 40 plays in all. A large part of his work is concerned with the social, literary, and philosophical life of Athens itself and with themes provoked by the great Peloponnesian War (431–404 BC). This war was essentially a conflict between imperialist Athens and conservative Sparta and so was long the dominant issue in Athenian politics. Aristophanes was naturally an opponent of the more or less bellicose statesmen who controlled the government of Athens throughout the better part of his maturity. Aristophanes lived to see the revival of Athens after its defeat by Sparta. He died in about 388 BC.

Importance of Peloponnesian War on his work

DRAMATIC AND LITERARY ACHIEVEMENTS

Aristophanes' reputation has stood the test of time; his plays have been frequently produced on the 20th-century stage in numerous translations, which manage with varying degrees of success to convey the flavour of Aristophanes' puns, witticisms, and topical allusions. But it is not easy to say why his comedies still appeal to an audience almost 2,500 years after they were written. In the matter of plot construction Aristophanes' comedies are often loosely put together, are full of strangely inconsequential episodes, and often degenerate at their end into a series of disconnected and boisterous episodes. Aristophanes' greatness lies in the wittiness of his dialogue; in his generally good-humoured though occasionally malevolent satire; in the brilliance of his parody, especially when he mocks the controversial tragedian Euripides; in the ingenuity and inventiveness, not to say the laughable absurdity, of his comic scenes born of imaginative fantasy; in the peculiar charm of his choric songs, whose freshness can still be conveyed in languages other than Greek; and, at least for audiences of a permissive age, in the licentious frankness of many scenes and allusions in his comedies.

THE PLAYS

Babylonians. This comedy, which is extant only in fragments, was produced at the festival of the Great Dionysia. The festival was attended by delegates of the city-states, which were theoretically "allies" but were in practice satellites of Athens. Because *Babylonians* (426 BC; Greek *Babylōnioi*) not only virulently attacked Cleon, the demagogue then in power in Athens, but also showed the "allies" as the slaves of the Athenian Demos (a personification of the Athenian citizen electorate), Aristophanes was impeached by Cleon. Though the details are not known, he seems to have been let off lightly.

Acharnians. This is the earliest of the 11 comedies of Aristophanes that have survived intact. *Acharnians* (425 BC; Greek *Acharneis*) is a forthright attack on the folly of the war. Its farmer-hero, Dicaeopolis, is tired of the Peloponnesian War and therefore secures a private peace treaty with the Spartans for himself in spite of the violent opposition of a chorus of embittered and bellicose old charcoal burners of Acharnae. Dicaeopolis takes advantage of his private treaty to trade with the allies of the Spartans. The Athenian commander Lamachus tries to stop him, but by the end of the play Lamachus slumps wounded and dejected while Dicaeopolis enjoys a peacetime life of food, wine, and sex.

Knights. This play shows how little Aristophanes was affected by the prosecution he had incurred for *Babylonians. Knights* (424 BC; Greek *Hippeis*) consists of a violent attack on the same demagogue, Cleon, who is depicted as the favourite slave of the stupid and irascible Demos until he is, at last, ousted from his position of influence and authority by Agoracritus, a sausage seller who is even more scoundrelly and impudent than Cleon.

Clouds. This play (423 BC; Greek *Nephelai*) is an attack on "modern" education and morals as imparted and taught by the radical intellectuals known as the Sophists. The main victim of the play is the leading Athenian thinker and teacher Socrates, who is purposely (and unfairly) given many of the standard characteristics of the Sophists. In the play Socrates is consulted by an old rogue, Strepsiades ("Twisterson"), who wants to evade his debts. The instruction at Socrates' academy, the Phrontisterion ("Thinking Shop"), which consists of making a wrong argument sound right, enables Strepsiades' son to defend the beating of his own father. At the play's end the Phrontisterion is burned to the ground.

Attack on Sophists

Wasps. This comedy satirized the litigiousness of the Athenians in the person of the mean and waspish old man Philocleon ("Love-Cleon"), who has a passion for serving on juries. In *Wasps* (422 BC; Greek *Sphēkes*) Philocleon's son, Bdelycleon ("Loathe-Cleon"), arranges for his father to hold a "court" at home; but, since the first "case" to be heard is that of the house dog accused of the theft of a cheese, Philocleon is finally cured of his passion for the law courts and instead becomes a boastful and uproarious drunkard. The play's main political target is the exploitation by Cleon of the Athenian system of large subsidized juries.

Peace. This play was staged seven months or so after both Cleon and Brasidas, the two main champions of the war policy on the Athenian and Spartan sides respectively, had been killed in battle and, indeed, only a few weeks before the ratification of the Peace of Nicias (? March 421 BC), which suspended hostilities between Athens and Sparta for six uneasy years. In *Peace* (421 BC; Greek *Eirēnē*) the war-weary farmer Trygaeus ("Vintager") flies to heaven on a monstrous dung beetle to find the lost goddess Peace, only to discover that the God of War has buried Peace in a pit. With the help of a chorus of farmers Trygaeus rescues her, and the play ends with a joyful celebration of marriage and fertility.

Birds. This play can be regarded merely as a "comedy of fantasy," but some scholars see *Birds* (414 BC; Greek *Ornithes*) as a political satire on the imperialistic dreams that had led the Athenians to undertake their ill-fated expedition of 415 BC to conquer Syracuse in Sicily. Peisthetaerus ("Trusty") is so disgusted with his city's bureaucracy that he persuades the birds to join him in building a new city that will be suspended in between heaven and earth; it is named Nephelokokkygia and is the original Cloudcuckooland. The city is built, and Peisthetaerus and his bird comrades must then fend off the undesirable humans who want to join them in their new Utopia. He and the birds finally even starve the Olympian gods into cooperating with them. *Birds* is Aristophanes' most fantastical play, but its escapist mood possibly echoes the dramatist's sense of Athens' impending decline.

Lysistrata. This comedy was written not long after the catastrophic defeat of the Athenian expedition to Sicily (413 BC) and not long before the revolt of the Four Hundred in Athens, whereby an oligarchic regime ready to make peace with Sparta was set up (411 BC). *Lysistrata* (411 BC; Greek *Lysistratē*) depicts the seizure of the Acropolis and of the treasury of Athens by the city's women who, at Lysistrata's instigation, have, together with all the women of Greece, declared a sex strike until such time as the men will make peace. The women defy their menfolk until the peace is arranged, after which both the Athenian and Spartan wives are reunited with their husbands. The play is a strange mixture of humour, indecency, gravity, and farce.

Plot of *Lysistrata*

Women at the Thesmophoria. In *Women at the Thesmophoria* (411 BC; Greek *Thesmophoriazousai*) Euripides has discovered that the women of Athens, angered by his constant attacks upon them in his tragedies, mean to

discuss during their coming festival (the Thesmophoria) the question of contriving his death. Euripides tries to persuade the effeminate Agathon, a tragic poet, to plead his cause. Agathon refuses, and Euripides persuades his brother-in-law Mnesilochus to undertake the assignment. Mnesilochus is disguised with great thoroughness as a woman and sent on his mission, but his true sex is discovered and he is at once seized by the women. There follow three scenes in which he tries unsuccessfully to escape; all three involve brilliant parodies of Euripides' tragedies, and all three attempts fail. Finally, Euripides himself arrives and succeeds in rescuing his advocate by promising never again to revile women.

Frogs. This is a literary comedy. In *Frogs* (405 BC; Greek *Batrachoi*) Dionysus, the god of drama, is concerned about the poor quality of present-day tragedy in Athens now that his recent favourite, Euripides, is dead. Dionysus disguises himself as the hero Heracles and goes down to Hades to bring Euripides back to the land of the living. As the result, however, of a competition arranged between Euripides and his great predecessor, Aeschylus, Dionysus is won over to the latter's cause and returns to earth with Aeschylus, instead, as the one more likely to help Athens in its troubles.

Women at the Ecclesia. In *Women at the Ecclesia* (*c.* 392 BC; Greek *Ekklēsiazousai*) the women of Athens dress up as men, take over the Ecclesia (the Athenian democratic assembly), and introduce a communistic system of wealth, sex, and property. It is not one of Aristophanes' more appealing plays.

Wealth. The last of the author's plays to be performed in his lifetime, *Wealth* (388 BC; Greek *Ploutos*) is a somewhat moralizing work and does not enhance his reputation—though, as suggested, it may have inaugurated the Middle Comedy.

Shortly after producing his *Wealth,* Aristophanes died, leaving two plays (now lost), the *Aiolosikon* and the *Kokolos,* which his son staged *c.* 387 BC; both of them are generally assumed to have been mythological burlesques.

(M.Pl./O.T.)

MAJOR WORKS

Aeschylus

WORKS: *Persai* (472 BC; *Persians*); *Hepta epi Thēbais* (467 BC; *Seven Against Thebes*); *Hiketides* (*c.* 463 BC; Latin trans., *Supplices;* Eng. trans., *Suppliants*); the trilogy known as the *Oresteia* (458 BC), comprising *Agamemnōn, Choēphoroi* (*Libation Bearers*), and *Eumenides*; *Promētheus desmōtēs* (date uncertain, probably late; *Prometheus Bound*).

TEXTS: The Greek text is available in Denys Page (ed.), *Aeschyli septem quae supersunt tragoedias* (1972, reprinted 1975), in the Oxford Classical Text series. Herbert Weir Smyth (trans.), *Aeschylus,* 2 vol. (1922–26, reprinted 1973–83), in the Loeb Classical Library series, presents the original text with an English translation.

RECOMMENDED EDITIONS: *Aeschylus,* 2 vol. (1953–56, reissued 1967), contains English translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. Recent editions of the trilogy are *Oresteia,* trans. by Hugh Lloyd-Jones, 3 vol. (1979, reissued in 1 vol., 1982); and *The Oresteia,* trans. by Robert Fagles (1975, reissued 1984).

Sophocles

WORKS: *Aias mastigophoros* (probably before 441 BC; *Ajax*); *Antigonē* (*c.* 442–441 BC); *Oidipous Tyrannos* (soon after 430 BC; Latin trans., *Oedipus Rex;* Eng. trans., *Oedipus the King*); *Trachiniai* (possibly after 430 BC; *Trachinian Women*); *Ēlektra* (between 418 and 410 BC; *Electra*); *Philoktētēs* (409 BC; *Philoctetes*); *Oidipous epi Kolōnō* (produced posthumously in 401 BC; *Oedipus at Colonus*); *Ichneutai* (sizable fragments; date uncertain; *Trackers*).

TEXTS: The Greek text is available in A.C. Pearson (ed.), *Sophoclis fabulae* (1924, reprinted 1975), in the Oxford Classical Text series; and R.D. Dawe (ed.), *Sophoclis tragoediae,* vol. 1 and 2 (1975–79), part of the Teubner series. Greek text with English translation is presented in R.C. Jebb (ed. and trans.), *Sophocles,* 7 vol. (1883–96, reprinted from various editions, 1976).

RECOMMENDED EDITIONS: *Sophocles,* 2 vol. (1954–69), contains translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. See also Robert Fagles (trans.), *The Three Theban Plays* (1982, reissued 1984), which includes *Antigonē, Oedipus the King,* and *Oedipus at Colonus.*

Euripides

WORKS: *Alkēstis* (438 BC; *Alcestis*); *Mēdeia* (431 BC; *Medea*); *Hērakleidai* (*c.* 430 BC; *Children of Heracles*); *Hippolytos* (428 BC; *Hippolytus*); *Andromachē* (*c.* 426 BC); *Hekabē* (*c.* 425 BC; *Hecuba*); *Hiketides* (*c.* 423 BC; Latin trans., *Supplices;* Eng. trans., *Suppliants*); *Ēlektra* (*c.* 418 BC; *Electra*); *Hēraklēs mainomenos* (*c.* 416 BC; Latin trans., *Hercules furens;* Eng. trans., *Madness of Heracles*); *Trōades* (415 BC; *Trojan Women*); *Iōn* (*c.* 413 BC); *Iphigeneia en Taurois* (*c.* 413 BC; Latin trans., *Iphigenia in Tauris;* Eng. trans., *Iphigenia Among the Taurians*); *Helenē* (412 BC; *Helen*); *Phoinissai* (*c.* 409 BC; *Phoenician Women*); *Orestēs* (408 BC); *Iphigeneia en Aulidi* (*c.* 406 BC; *Iphigenia at Aulis*); *Bakchai* (*c.* 406 BC; Latin trans., *Bacchae;* Eng. trans., *Bacchants*); *Kyklōps* (date unknown; *Cyclops*); *Rhēsos* (authorship disputed; date unknown; *Rhesus*); *Hypsipylē* (sizable fragments; date uncertain).

TEXTS: The Greek text is available in *Euripidis fabulae,* ed. by J. Diggle (1981–), a new Oxford Classical Text replacing the 3-vol. text ed. by Gilbert Murray, 1902–09; 2 vol. of the Diggle work have appeared to 1986.

RECOMMENDED EDITIONS: *Euripides,* 5 vol. (1955–59), contains English translations of all the plays and is part of The Complete Greek Tragedies series, ed. by Richmond Lattimore and David Grene. All the complete plays excluding *Cyclops* and *Rhesus* are found in the Penguin Classics series, trans. by Philip Vellacott, *Three Plays* (1953, reissued 1974), *Medea, and Other Plays* (1963), *Orestes, and Other Plays* (1972), and *The Bacchae, and Other Plays,* rev. ed. (1972). See also Geoffrey S. Kirk (trans.), *The Bacchae* (1970, reprinted 1979 as *The Bacchae of Euripides*).

Aristophanes

WORKS: *Babylōnioi* (426 BC; *Babylonians*); *Acharneis* (425 BC; *Acharnians*); *Hippeis* (424 BC; *Knights*); *Nephelai* (423 BC; *Clouds*); *Sphēkes* (422 BC; *Wasps*); *Eirēnē* (421 BC; *Peace*); *Ornithes* (414 BC; *Birds*); *Lysistratē* (411 BC; *Lysistrata*); *Thesmophoriazousai* (411 BC; *Women at the Thesmophoria*); *Batrachoi* (405 BC; *Frogs*); *Ekklēsiazousai* (*c.* 392 BC; *Women at the Ecclesia*); *Ploutos* (388 BC; *Wealth*).

TEXTS: The Greek text is available in Victor Coulon (ed.) and Hilaire Van Daele (trans.), *Aristophane,* 5 vol. (1923–30), part of the Budé series. Greek text with English translation is presented in Benjamin Bickley Rogers (ed. and trans.), *Aristophanous komodiai: The Comedies of Aristophanes,* 6 vol. in 7 (1902–16; reissued as part of the Loeb Classical Library series, 3 vol., 1924, reissued 1979–82); and in Alan H. Sommerstein (ed. and trans.), *The Comedies of Aristophanes* (1980–), five plays in single volumes having appeared to 1986.

RECOMMENDED EDITIONS: All the plays appear in English translation in three separate volumes in the Penguin Classics series, trans. respectively by David Barrett (1964, reprinted 1976), by Alan H. Sommerstein (1973, reprinted 1977), and by David Barrett and Alan H. Sommerstein (1978).

BIBLIOGRAPHY

Greek tragedy in general: H.D.F. KITTO, *Greek Tragedy: A Literary Study,* 3rd ed. (1961, reissued 1976), a lively survey, but becoming dated; RICHMOND LATTIMORE, *Story Patterns in Greek Tragedy* (1964, reissued 1969); ARTHUR PICKARD-CAMBRIDGE, *The Dramatic Festivals of Athens,* 2nd ed. rev. by JOHN GOULD and D.M. LEWIS (1968), a standard work on the practical arrangements; H.C. BALDRY, *The Greek Tragic Theatre* (1971), a simple, orthodox introduction; ALBIN LESKY, *Greek Tragic Poetry* (1983; originally published in German, 3rd rev. ed., 1972); ERIKA SIMON, *The Ancient Theatre* (1982; originally published in German, 2nd ed., 1981), a concise and expert introduction to staging; JEAN-PIERRE VERNANT and PIERRE VIDAL-NAQUET, *Tragedy and Myth in Ancient Greece* (1981; originally published in French, 1972), stimulating structuralist essays; BRIAN VICKERS, *Towards Greek Tragedy: Drama, Myth, Society* (1973, reprinted 1979), long but thought-provoking; OLIVER TAPLIN, *Greek Tragedy in Action* (1978), emphasis on the significance of performance; BERNARD KNOX, *Word and Action: Essays on the Ancient Theater* (1979), a collection of important essays; DONALD J. MASTRONARDE, *Contact and Discontinuity: Some Conventions of Speech and Action on the Greek Tragic Stage* (1979), a specialist study of dialogue; R.G.A. BUXTON, *Persuasion in Greek Tragedy: A Study of Peitho* (1982); and ERICH SEGAL (ed.), *Greek Tragedy: Modern Essays in Criticism* (U.K. title, *Oxford Readings in Greek Tragedy,* 1983), a well-chosen and varied selection of articles.

Aeschylus: Studies of special topics relating to Aeschylus' plays include ANTHONY J. PODLECKI, *The Political Background of Aeschylean Tragedy* (1966); A.F. GARVIE, *Aeschylus' "Supplices": Play and Trilogy* (1969); GEORGE THOMSON, *Aeschylus and Athens: A Study in the Social Origins of Drama,* 4th ed. (1973), a Marxist study; ANNE LEBECK, *The Oresteia: A Study in Language and Structure* (1971), on the significant connections of imagery; R.P. WINNINGTON-INGRAM, *Studies in*

Aeschylus (1983), a collection of insightful essays; THOMAS G. ROSENMEYER, *The Art of Aeschylus* (1982), a critical study; D.J. CONACHER, *Aeschylus' "Prometheus Bound": A Literary Commentary* (1980); MARK GRIFFITH, *The Authenticity of "Prometheus Bound"* (1977), a powerful attack on authenticity; and OLIVER TAPLIN, *The Stagecraft of Aeschylus: The Dramatic Use of Exits and Entrances in Greek Tragedy* (1978), on dramatic techniques and meanings.

Sophocles: Influential interpretations of Sophocles' life and works are T.B.L. WEBSTER, *An Introduction to Sophocles,* 2nd ed. (1969, reprinted 1979); G.H. GELLIE, *Sophocles: A Reading* (1972); R.P. WINNINGTON-INGRAM, *Sophocles: An Interpretation* (1980), an important study; CHARLES SEGAL, *Tragedy and Civilization: An Interpretation of Sophocles* (1981), a structuralist approach; BERNARD M.W. KNOX, *The Heroic Temper: Studies in Sophoclean Tragedy* (1964, reprinted 1983), an important study of Sophocles' stubborn "heroes"; KARL REINHARDT, *Sophocles* (1979; originally published in German, 3rd ed., 1947), a great classic; G.M. KIRKWOOD, *A Study of Sophoclean Drama* (1958, reprinted 1967); R.W.B. BURTON, *The Chorus in Sophocles' Tragedies* (1980); and DAVID SEALE, *Vision and Stagecraft in Sophocles* (1982).

Euripides: Critical works include G.M.A. GRUBE, *The Drama of Euripides* (1941, reprinted 1973), a survey; D.J. CONACHER, *Euripidean Drama: Myth, Theme and Structure* (1967), a helpful play-by-play survey, including a useful bibliography; GILBERT MURRAY, *Euripides and His Age,* 2nd ed. (1946, reissued 1979), an idealistic introduction; T.B.L. WEBSTER, *The Tragedies of Euripides* (1967), helpful on the lost tragedies; R.P. WINNINGTON-INGRAM, *Euripides and Dionysus: An Interpretation of the Bacchae* (1948, reprinted 1969), an enterprising study; WILLIAM RITCHIE, *The Authenticity of the Rhesus of Euripides* (1964), a scholarly though unsuccessful case for authorship; ANNE PIPPIN BURNETT, *Catastrophe Survived: Euripides Plays of Mixed Reversal* (1971, reprinted 1985), an original interpretation of seven plays; PIETRO PUCCI, *The Violence of Pity in Euripides' "Medea"* (1980); SHIRLEY A. BARLOW, *The Imagery of Euripides: A Study in the Dramatic Use of Pictorial Language* (1971, reprinted 1974); and HELENE P. FOLEY, *Ritual Irony: Poetry and Sacrifices in Euripides* (1985), a structuralist study of four plays.

Aristophanes: Critical studies include GILBERT MURRAY, *Aristophanes* (1933, reprinted 1965), still lively; VICTOR EHRENBERG, *The People of Aristophanes: A Sociology of Old Attic Comedy,* 2nd ed. (1951, reprinted with revisions 1974); CEDRIC H. WHITMAN, *Aristophanes and the Comic Hero* (1964), concentrating on the central figure of each play; F.H. SANDBACH, *The Comic Theatre of Greece and Rome* (1977), especially the first chapters on Old Comedy; CARROLL MOULTON, *Aristophanic Poetry* (1981), artistic unity through poetry and imagery; ROSEMARY M. HARRIOTT, *Aristophanes: Poet & Dramatist* (1986), a study of sophistication of literary technique; K.J. DOVER, *Aristophanic Comedy* (1972), an introduction; KENNETH MCLEISH, *The Theatre of Aristophanes* (1980), on the mechanics of raising laughs; and JEFFREY HENDERSON (ed.), *Aristophanes: Essays in Interpretation* (1980).

(A.J.P./T.W./H.D.F.K./M.Pl./O.T.)

Greek Literature

Greek literature has a continuous history extending from the 1st millennium BC to the present day. From the beginning its writers were Greeks living not only in Greece proper but also in Asia Minor, the Aegean Islands, and Magna Graecia (Sicily and southern Italy). Later, after the conquests of Alexander the Great, Greek became the common language of the eastern Mediterranean lands and then of the Byzantine Empire. Literature in Greek was produced not only over a much wider area but also by those whose mother tongue was not Greek. Even before the Turkish conquest (1453) the area had begun to shrink again, and now it is chiefly confined to the territory of the republic of Greece.

This article is divided into the following sections:

Ancient Greek literature 352
- Stylistic periods 352
  - Preclassical period, to the end of the 6th century BC
  - Classical period, 5th and 4th centuries BC
  - Hellenistic and Greco-Roman periods
- The genres 353
  - Epic narrative
  - Lyric poetry
  - Tragedy
  - Comedy
  - History
  - Rhetoric and oratory
  - Philosophical prose
  - Late forms of poetry
  - Late forms of prose

Byzantine literature 357
- General characteristics 357
- Principal forms of writing 357
  - Nonliturgical poetry
  - Liturgical poetry
  - Historical works
  - Rhetoric

Modern Greek literature (after 1453) 358
- Post-Byzantine period 358
- Independence and after 359
  - Old Athenian School
  - Heptanesian School
  - Demoticism and folklorism, 1880–1922
  - Literature from 1922

Bibliography 360

Ancient Greek literature

Of the literature of ancient Greece only a relatively small proportion survives. Yet it remains important, not only because much of it is of supreme quality but also because until the mid-19th century the greater part of the literature of the Western world was produced by writers who were familiar with the Greek tradition, either directly or through the medium of Latin, who were conscious that the forms they used were mostly of Greek invention, and who took for granted in their readers some familiarity with classical literature.

STYLISTIC PERIODS

The history of ancient Greek literature may be divided into three periods: preclassical (to the end of the 6th century BC); classical (5th and 4th centuries BC); and Hellenistic and Greco-Roman (3rd century BC onward).

Preclassical period, to the end of the 6th century BC. The Greeks created poetry before they made use of writing for literary purposes, and from the beginning their poetry was intended to be sung or recited. (The art of writing was little known before the 7th century BC. The script used in Crete and Mycenae [Linear B] is not known to have been employed for other than administrative purposes, and after the destruction of the Mycenaean cities it was forgotten.) Its subject was myth—part legend, based sometimes on the dim memory of historical events; part folktale; and part primitive religious speculation. But since the myths were not associated with any religious dogma, even though they often treated of gods and heroic mortals, they were not authoritative and could be varied by a poet to express new concepts.

Thus, at an early stage Greek thought was advanced as poets refashioned their materials; and to this stage of preclassical literature belonged the epics ascribed to Homer, the *Iliad* and the *Odyssey,* retelling intermingled history and myth of the Mycenaean Age. These two great poems, standing at the beginning of Greek literature, established most of the literary conventions of the epic poem. The didactic poetry of Hesiod (*c.* 700 BC) was probably later in composition than Homer's epics and, though different in theme and treatment, continued the epic tradition.

Types of early Greek poetry

The several types of Greek lyric poetry originated in the preclassical period among the poets of the Aegean Islands and of Ionia on the coast of Asia Minor. Archilochus of Paros, of the 7th century BC, was the earliest Greek poet to employ the forms of elegy (in which the epic verse line alternated with a shorter line) and of personal lyric poetry. His work was very highly rated by the ancient Greeks but survives only in fragments; its forms and metrical patterns—the elegiac couplet and a variety of lyric metres—were taken up by a succession of Ionian poets. At the beginning of the 6th century Alcaeus and Sappho, composing in the Aeolic dialect of Lesbos, produced lyric poetry mostly in the metres named after them (the alcaic and the sapphic), which Horace was later to adapt to Latin poetry. No other poets of ancient Greece entered into so close a personal relationship with the reader as Alcaeus, Sappho, and Archilochus do. They were succeeded by Anacreon of Teos, in Ionia, who, like Archilochus, composed his lyrics in the Ionic dialect. Choral lyric, with musical accompaniment, belonged to the Dorian tradition and its dialect, and its representative poets in the period were Alcman in Sparta and Stesichorus in Sicily.

Both tragedy and comedy had their origins in Greece. "Tragic" choruses are said to have existed in Dorian Greece around 600 BC, and in a rudimentary dramatic form tragedy became part of the most famous of the Dionysian festivals, the Great, or City, Dionysia at Athens, about 534. Comedy, too, originated partly in Dorian Greece and developed in Attica, where it was officially recognized rather later than tragedy. Both were connected with the worship of Dionysus, god of fruitfulness and of wine and ecstasy.

Written codes of law were the earliest form of prose and were appearing by the end of the 7th century, when knowledge of reading and writing was becoming more widespread. No prose writer is known earlier than Pherecydes (*c.* 550 BC) of Syros, who wrote about the beginnings of the world; but the earliest considerable author was Hecataeus of Miletus, who wrote about both the mythical past and the geography of the Mediterranean and surrounding lands. To Aesop, a semi-historical, semi-mythological character of the mid-6th century, have been attributed the moralizing beast fables copied by later writers.

Classical period, 5th and 4th centuries BC. True tragedy was created by Aeschylus and reached its culmination with Sophocles and Euripides in the second half of the 5th century. Aristophanes, the greatest of the comedic poets, lived on into the 4th century, but the Old Comedy did not survive the fall of Athens in 404. (These four dramatists and their works are treated in detail in GREEK DRAMATISTS, THE CLASSICAL.)

The sublime themes of Aeschylean tragedy, in which man stands answerable to the gods and receives awe-inspiring

insight into their purposes, are exemplified in the three plays of the *Oresteia*. The tragedy of Sophocles made a progress toward both dramatic complexity and naturalness while remaining orthodox in its treatment of religious and moral issues. Euripides handled his themes on the plane of skeptical enlightenment and doubted the traditional picture of the gods. Corresponding development of dramatic realization accompanied the shift of vision: the number of individual actors was raised to three, each capable of taking several parts.

The three stages of dramatic comedy

The Old Comedy of Aristophanes was established later than tragedy but preserved more obvious traces of its origin in ritual; for the vigour, wit, and indecency with which it keenly satirized public issues and prominent persons clearly derived from the primitive ribaldry of the Dionysian festival. Aristophanes' last comedies show a transition, indicated by the dwindling importance of the chorus, toward the Middle Comedy, of which no plays are extant. This phase was followed toward the beginning of the 3rd century by the New Comedy, introduced by Menander, which turned for its subjects to the private fictional world of ordinary people. Later adaptations of New Comedy in Latin by Plautus and Terence carried the influence of his work on to medieval and modern times.

In the 5th century, Pindar, the greatest of the Greek choral lyrists, stood outside the main Ionic-Attic stream and embodied in his splendid odes a vision of the world seen in terms of aristocratic values that were already growing obsolete. Greek prose came to maturity in this period. Earlier writers such as Anaxagoras the philosopher and Protagoras the Sophist used the traditional Ionic dialect, as did Herodotus the historian. His successors in history, Thucydides and Xenophon, wrote in Attic.

The works of Plato and Aristotle, of the 4th century, are the most important of all the products of Greek culture in the intellectual history of the West. They were preoccupied with ethics, metaphysics, and politics as man's highest study and, in the case of Aristotle, extended the range to include physics, natural history, psychology, and literary criticism. They have formed the basis of Western philosophy and, indeed, they determined, for centuries to come, the development of European thought. (For detailed treatment see ARISTOTELIANISM, ARISTOTLE AND and PLATONISM, PLATO AND.)

This was also a golden age for rhetoric and oratory, first taught by Corax of Syracuse in the 5th century. The study of rhetoric and oratory raised questions of truth and morality in argument, and thus it was of concern to the philosopher as well as to the advocate and the politician and was expounded by teachers, among whom Isocrates was outstanding. The orations of Demosthenes, a statesman of 4th-century Athens and the most famous of Greek orators, are preeminent for force and power.

Hellenistic and Greco-Roman periods. In the huge empire of Alexander the Great, Macedonians and Greeks composed the new governing class; and Greek became the language of administration, a new composite dialect based to some extent on Attic and called the Koine, or common language. Everywhere the traditional city-state was in decline, and the individual was becoming aware of his isolation and seeking consolidation and satisfaction outside corporate society. Artistic creation now came under private patronage, and, except for Athenian comedy, compositions were intended for a small, select audience that admired polish, erudition, and subtlety.

Importance of the Museum at Alexandria

An event of great importance for the development of new tendencies was the founding of the Museum, the shrine of the Muses with its enormous library, at Alexandria. The chief librarian was sometimes a poet as well as tutor of the heir apparent. The task of accumulating and preserving knowledge begun by the Sophists and continued by Aristotle and his adherents was for the first time properly endowed. Through the researches of the Alexandrian scholars, texts of ancient authors were preserved.

The Hellenistic period lasted from the end of the 4th to the end of the 1st century BC. For the next three centuries, until Constantinople became the capital of the Byzantine Empire, Greek writers were conscious of belonging to a world of which Rome was the centre.

THE GENRES

Epic narrative. At the beginning of Greek literature stand the two great epics, the *Iliad* and the *Odyssey*. Some features of the poems reach far into the Mycenaean age, perhaps to 1500 BC, but the written works are traditionally ascribed to Homer; in something like their present form they probably date to the 8th century. (Homer and the works attributed to him are treated in detail in HOMERIC EPICS, THE.)

The *Iliad* and the *Odyssey*

The *Iliad* and the *Odyssey* are primary examples of the epic narrative, which in antiquity was a long narrative poem, in an elevated style, celebrating heroic achievement. The *Iliad* is the tragic story of the wrath of Achilles, son of a goddess and richly endowed with all the qualities that make men admirable. With his readiness to sacrifice all to honour, Achilles embodies the Greek heroic ideal; and the contrast between his superb qualities and his short and troubled life reflects the sense of tragedy always prevalent in Greek thought. Whereas the *Iliad* is tragedy, the *Odyssey* is tragicomedy. It is an enriched version of the old folktale of the wanderer's return and of his triumph over those who were usurping his rights and importuning his wife at home. Odysseus too represents a Greek ideal. Though by no means inadequate in battle, he works mainly by craft and guile; and it is by mental superiority that he survives and prevails.

Both poems were based on plots that grip the reader, and the story is told in language that is simple and direct, yet eloquent. The *Iliad* and the *Odyssey,* though they are the oldest European poetry, are by no means primitive. They marked the fulfillment rather than the beginning of the literary form to which they belong. They were essentially oral poems, handed down, developed, and added to over a vast period of time, a theme upon which successive nameless poets freely improvised. The world they reflect is full of inconsistencies; weapons belong to both the Bronze and Iron Ages, and objects of the Mycenaean period jostle others from a time five centuries later. Certain mysteries remain: the date of the great poet or poets who gave structure and shape to the two epics; the social function of poems that take several days to recite; and the manner in which these poems came to be recorded in writing probably in the course of the 6th century BC.

In the ancient world the *Iliad* and the *Odyssey* stood in a class apart among epic poems. Of these, there were a large number known later as the epic cycle. They covered the whole story of the wars of Thebes and Troy as well as other famous myths. A number of shorter poems in epic style, the Homeric Hymns, are of considerable beauty.

Didactic poetry

Didactic poetry was not regarded by the Greeks as a form distinct from epic. Yet the poet Hesiod belonged to an altogether different world from Homer. He lived in Boeotia in central Greece about 700 BC. In his *Works and Days* he described the ways of peasant life and incidentally described the dreary Boeotian plain afflicted by heat, cold, and the oppression of a "gift-devouring" aristocracy. He believed passionately in a Zeus who cared about right and wrong and in Justice as Zeus's daughter. Hesiod's other surviving poem, the *Theogony,* attempts a systematic genealogy of the gods and recounts many myths associated with their part in the creation of the universe. Middle Eastern influence is clearly to be seen, especially in some of the cruder speculation about the origin of the universe.

By the end of the 6th century the epic tradition was a spent force until its revival in the Hellenistic period, and the few composers of epic narrative left little but their names.

Lyric poetry. Hesiod, unlike Homer, told something of himself, and the same is true of the lyric poets. Except for Pindar and Bacchylides at the end of the classical period, only fragments of the works of these poets survive. There had always been lyric poetry in Greece. All the great events of life as well as many occupations had their proper songs, and here too the way was open to advance from the anonymous to the individual poet.

The word "lyric" covers many sorts of poem. On the one hand, poems sung by individuals or chorus to the lyre, or sometimes to the flute, were called melic; elegiacs, in which the epic hexameter, or verse line of six metri-

cal feet, alternated with a shorter line, were traditionally associated with lamentation and a flute accompaniment; but they were also used for personal poetry, spoken as well as sung. Iambics (verse of iambs, or metrical units, basically of four alternately short and long syllables) were the verse form of the lampoon. Usually of an abusive or satirical—burlesque and parodying—character, they were not normally sung.

If Archilochus of Paros in fact was writing as early as 700 BC, he was the first of the post-epic poets. The fragments reflect the turbulent life of an embittered adventurer. Scorn both of men and of convention is the emotion that seems uppermost, and Archilochus was possessed of tremendous powers of invective. Of lesser stature than Archilochus were his successors, Semonides (often mistakenly identified with Simonides) of Amorgos and Hipponax of Ephesus.

Elegiac poets

Like the iambic writers, the elegiac poets came mostly from the islands and the Ionian regions of Asia Minor. Chief among them were Callinus of Ephesus and Mimnermus of Colophon. On the mainland of Greece, Tyrtaeus roused the spirit of the Spartans in their desperate struggle with the Messenian rebels in the years after 650. His martial poems are perhaps of more historical than literary interest. The same is to some extent true of the poems in elegiac, iambic, and trochaic (the latter a metre basically of four alternately long and short syllables) metres by Solon, an Athenian statesman, who used his poetry as a vehicle for propaganda. Xenophanes (born about 560 BC) rather in the same way used his poems to propagate his revolutionary religious and ethical ideas. The elegiacs attributed to Theognis seem to be poems of various dates suitable for use at drinking parties. Many of them were actually by Theognis himself (about 540 BC). Some give uninhibited expression to his hatred of the lower class rulers who had ousted the aristocracy of Megara; others are love poems to the boy Cyrnus; still others are gnomic commonplaces of Greek wisdom and morality.

About the beginning of the 6th century a new kind of poetry made its appearance in the island of Lesbos. It was composed in the local Aeolic dialect by members of the turbulent and factious aristocracy. Alcaeus (born about 620 BC), absorbed in political feuds and in civil war, expressed with striking directness searing hate and blind exultation. With the same directness and infinite grace, Sappho, a younger contemporary who seems to have enjoyed a freedom unknown to the women of mainland Greece, told of her loves and hates, though little is known of her relations with the girls named in her poems. The surviving works by their successor in personal lyric, Anacreon of Teos, suggest a more convivial amorousness.

Choral lyric was associated with the Dorian parts of the Greek mainland and the settlements in Sicily and south Italy, whereas poetry for solo performance was a product of the Ionian coast and the Aegean Islands. Thus choral song came to be conventionally written in a Doric dialect.

Choral lyric, which had lyre and flute accompaniments, was highly complicated in structure. It did not use traditional lines or stanzas; but the metre was formed afresh for each poem and never used again in exactly the same form, though the metrical units from which the stanzas, or strophes, were built up were drawn from a common stock and the form of the strophe was usually related to the accompanying dance. This elaborate art form was connected mainly with the cult of the gods or, as in the case of Pindar, the celebration of the victors in the great Hellenic games.

The earliest poet of choral lyrics of whose work anything has survived was Alcman of Sparta (about 620 BC). Somewhat later Stesichorus worked in Sicily, and his lyric versions of the great myths marked an important stage in the development of these stories. Simonides of Ceos, in Ionia, was among the most versatile of Greek poets. He was famed for his pathos, but today he is best known for his elegiac epitaphs, especially those on the Greek soldiers who fell in the struggle against Persia.

Pindar's odes

The supreme poet of choral lyric was Pindar from Thebes in Boeotia (born 518 or possibly 522–died after 446 BC), who is known mainly by his odes in honour of the victors at the great games held at Olympia, Delphi, the Isthmus of Corinth, and Nemea. The last of the lyric poets was Bacchylides (flourished 5th century BC), whose work, though often exquisite, is empty, reflecting the declining significance of myth.

Tragedy. Tragedy may have developed from the dithyramb, the choral cult song of the god Dionysus. Arion of Lesbos, who is said to have worked at Corinth in about 600, is credited with being the first to write serious poetry in this medium. Thespis (6th century BC), possibly combining with dithyrambs something of the Attic ritual of Dionysus of Eleutherae, is credited with having invented tragedy by introducing an actor who conversed with the leader of the chorus. These performances became a regular feature of the great festival of Dionysus at Athens about 534 BC. Aeschylus introduced a second actor, though his drama was still centred in the chorus, to whom, rather than to each other, his actors directed themselves.

At the tragic contests at the Dionysia each of three competing poets produced three tragedies and a satyr play, or burlesque, in which there was a chorus of satyrs. Aeschylus, unlike later poets, usually made of his three tragedies a dramatic whole, treating a single story, as in the *Oresteia,* the only complete trilogy that has survived. His main concern was not dramatic excitement and the portrayal of character but rather the presentation of human action in relation to the overriding purpose of the gods.

The tragedies of Aeschylus, Sophocles, and Euripides

His successor was Sophocles, who abandoned for the most part the practice of writing in trilogies, reduced the importance of the chorus, and introduced a third actor. His work too was based on myth, but whereas Aeschylus tried to make more intelligible the working of the divine purpose in its effects on man's life, Sophocles was readier to accept the gods as given and to reveal the values of life as it can be lived within the traditional framework of moral standards. Sophocles' skill in control of dramatic movement and his mastery of speech were devoted to the presentation of the decisive, usually tragic, hours in the lives of men and women at once "heroic" and human, such as Oedipus.

Euripides, last of the three great tragic poets, belonged to a different world. When he came to manhood, traditional beliefs were scrutinized in the light of what was claimed by Sophist philosophers, not always unjustifiably, to be reason; and this was a test to which much of Greek religion was highly vulnerable. The whole structure of society and its values was called into question. This movement of largely destructive criticism was clearly not uncongenial to Euripides. But as a dramatic poet he was bound to draw his material from myths, which, for him, had to a great extent lost their meaning. He adapted them to make room for contemporary problems, which were his real interest. Many of his plays suffer from a certain internal disharmony, yet his sensibilities and his moments of psychological insight bring him far closer than most Greek writers to modern taste. There are studies, wonderfully sympathetic, of wholly unsympathetic actions in the *Medea* and *Hippolytus;* a vivid presentation of the beauty and horror of religious ecstasy in the *Bacchants;* in the *Electra,* a reduction to absurdity of the values of a myth that justifies matricide; in *Helen* and *Iphigenia Among the Taurians,* melodrama with a faint flavour of romance.

Comedy. Like tragedy, comedy arose from a ritual in honour of Dionysus, in this case full of abuse and obscenity connected with averting evil and encouraging fertility. The parabasis, the part of the play in which the chorus broke off the action and commented on topical events and characters, was probably a direct descendant of such revels. The dramatic element may have been derived from the secular Dorian comedy without chorus, said to have arisen at Megara, which was developed at Syracuse by Epicharmus (*c.* 530–*c.* 440). Akin to this kind of comedy seems to have been the mime, a short realistic sketch of scenes from everyday life. These were written rather later by Sophron of Syracuse; only fragments have survived but they were important for their influence on Plato's dialogue form and on Hellenistic mime. At Athens, comedy became an official part of the celebrations of Dionysus in 486 BC. The first great comic poet was Cratinus. About 50 years

later Aristophanes and Eupolis refined somewhat the wild robustness of the older poet. But even so, for boldness of fantasy, for merciless invective, for unabashed indecency, and for freedom of political criticism, there is nothing like the Old Comedy of Aristophanes, whose work alone has survived. Cleon the politician, Socrates the philosopher, Euripides the poet were alike the victims of his masterly unfairness, the first in *Knights;* the second in *Clouds;* and the third in *Women at the Thesmophoria* and *Frogs;* whereas in *Birds* the Athenian democracy itself was held up to a kindlier ridicule. Aristophanes survived the fall of Athens in 404, but the Old Comedy had no place in the revived democracy.

The gradual change from Old to Middle Comedy took place in the early years of the 4th century. Of Middle Comedy, no fully developed specimen has survived. It seems to have been distinguished by the disappearance of the chorus and of outspoken political criticism and by the growth of social satire and of parody; Antiphanes and Alexis were the two most distinguished writers. The complicated plots in some of their plays led to the development of the New Comedy at the end of the century, which is best represented by Menander. One complete play, the *Dyscolus,* and appreciable fragments of others are extant on papyrus. New Comedy was derived in part from Euripidean tragedy; its characteristic plot was a translation into terms of city life of the story of the maiden—wronged by a god—who bears her child in secret, exposes it, and recognizes it years after by means of the trinkets she had put into its cradle.

History. The first great writer of history was Herodotus of Halicarnassus, who was also a geographer and anthropologist. The theme of his history, written in large part for Athenian readers, is the clash between Europe and Asia culminating in the Persian War. The account of the war itself, which occupies roughly the second half of the work, must have been composed by means of laborious inquiry from those whose memories were long enough to recall events that happened when Herodotus was a child or earlier. The whole history, though in places badly put together, is magnificent in its compass and unified by the consciousness of an overriding power keeping the universe and mankind in check.

Thucydides (*c.* 460–*c.* 400) was perhaps the first person to apply a first-class mind to a prolonged examination of the nature of political power and the factors by which policies of states are determined. As a member of the board of generals he acquired inside knowledge of the way policy is shaped. After his failure to save Amphipolis in 424, he spent 20 years in exile, which he used as an opportunity for getting at the truth from both sides. The result was a history of the war narrowly military and political but of the most penetrating quality. Thucydides investigated the effect on individuals and nations both of psychological characteristics and of chance. His findings were interpreted through the many speeches given to his characters.

The histories of Thucydides and Xenophon

Just as Thucydides had linked his work to the point at which Herodotus had stopped, so Xenophon (431–died before 350) began his *Hellenica* where Thucydides' unfinished history breaks off in 411. He carried his history down to 362. His work was superficial by comparison with that of Thucydides, but he wrote with authority of military affairs and appears at his best in the *Anabasis,* an account of his participation in the enterprise of the Greek mercenary army, with which the Persian prince Cyrus tried to expel his brother from the throne, and of the adventurous march of the Greeks, after the murder of their leaders by the Persians, from near Babylon to the Black Sea coast. Xenophon also wrote works in praise of Socrates, of whom his understanding was superficial.

No other historical writing of the 4th century has survived except for a substantial papyrus fragment containing a record of events of the years 396–395. Later, history declined to being merely a province of rhetoric.

Rhetoric and oratory. In few societies has the power of fluent and persuasive speech been more highly valued than it was in Greece, and even in Homer there are speeches that are pieces of finished rhetoric. But it was the rise of democratic forms of government that provided a great incentive to study and instruction in the arts of persuasion, which were equally necessary for political debate in the assembly and for attack and defense in the law courts.

Origin of the study of rhetoric

The formal study of rhetoric seems to have originated in Syracuse *c.* 460 BC with Corax and his pupils Tisias and Gorgias (died *c.* 376); Gorgias was influential also in Athens. Corax is reputed to have been the first to write a handbook on the art of rhetoric, dealing with such topics as arguments from probability and the parts into which speeches should be divided. Most of the Sophists had pretensions as teachers of the art of speaking, especially Protagoras, who postulated that the weaker of two arguments could by skill be made to prevail over the stronger, and Prodicus of Ceos.

Antiphon (*c.* 480–411), the first professional speech writer, was an influential opponent of democracy. Three speeches of his, all dealing with homicide cases, have been preserved, as have three "tetralogies," sets of two pairs of speeches containing the arguments to be used on both sides in imaginary cases of homicide. In them primitive ideas are expressed concerning bloodguilt and the duty of vengeance. Antiphon's style is bare and rather crudely antithetical. Gorgias from Sicily, who visited Athens in 427, introduced an elaborate balance and symmetry emphasized by rhyme and assonance. Thrasymachus of Chalcedon made a more solid contribution to the evolution of a periodic and rhythmical style.

Andocides (*c.* 440–died after 391), an orator who spent much of his life in exile from Athens, wrote three speeches containing vivid narrative; but as an orator he was admittedly amateurish. Lysias (*c.* 455–died after 380) lived at Athens for many years as a resident alien and supported himself by writing speeches when he lost his wealth. His speeches, some of them written for litigants of humble station, show dexterous adaptation to the character of the speaker, though the most interesting of all is his own attack on Eratosthenes, one of the Thirty Tyrants imposed on Athens by the Spartans in 404 BC.

Orators of the 4th century BC

The 12 extant speeches of Isaeus, who was active in the first half of the 4th century BC, throw light on aspects of Athenian law. Isocrates, who was influential in Athens for half a century before his death in 338, perfected a periodic prose style that, through the medium of Latin, was widely accepted as a pattern; and he helped give rhetoric its predominance in the educational system of the ancient world. In his writings, which took the form of speeches but were more like pamphlets, Isocrates shows some insight into the political troubles besetting Greece, with its endless bickering between cities incapable of cooperation.

The greatest of the orators was Demosthenes (384–322), supreme in vehemence and power, though lacking in some of the more delicate shadings of rhetorical skill. His speeches were mainly political, and he is best remembered for his energetic opposition to the rise of Macedonia under its king Philip II, embodied in the three "Philippics." After Demosthenes, oratory faded, together with the political setting to which it owed its preeminence. Two more 4th-century-BC writers need only be mentioned, Hyperides (*c.* 390–322) and Lycurgus (*c.* 390–324).

Philosophical prose. Prose as a medium of philosophy was written as early as the 6th century. Practitioners include Thales, Anaximander, Democritus, and Heracleitus. Philosophical prose was the greatest literary achievement of the 4th century. It was influenced by Socrates (who himself wrote nothing) and his characteristic method of teaching by question and answer, which led naturally to the dialogue. Alexamenus of Teos and Antisthenes, both disciples of Socrates, were the first to use it; but the greatest exponent of Socratic dialogue was the Athenian Plato (428/427–348/347). Shortly after Socrates' death in 399 Plato wrote some dialogues, mostly short; to this group of work belong the *Apology, Protagoras,* and *Gorgias.* In the decade after 385 he wrote a series of brilliant works, *Phaedo, Phaedrus, Symposium,* and the *Republic.* His Socrates is the most carefully drawn character in Greek literature. Subsequent dialogues became more austerely philosophical; Socrates tended increasingly to be a mere spokesman for Plato's thought; and in the last of his works, the *Laws,* he was replaced by a colourless "stranger."

Plato

Plato's style is a thing of matchless beauty, though ancient critics, who were likely to entangle themselves in the rules they had invented, found it too poetical.

Aristotle

Plato's pupil Aristotle (384–322) was admired in antiquity for his style; but his surviving works are all of the "esoteric" sort, intended for use in connection with his philosophical and scientific school, the Lyceum. They are without literary grace, and at times they approximate lecture notes. His works on literary subjects, the *Rhetoric,* and above all, the *Poetics,* had an immense effect on literary theory after the Renaissance. In the ancient world, Aristotelian doctrine was known mainly through the works of his successor Theophrastus (*c.* 372–288/287), now lost except for two books on plants and a famous collection of 30 *Characters,* sketches of human types much imitated by English writers of the 17th century.

With Theophrastus, Attic prose died out for a time. Technical prose in this period was produced in abundance; and rhetoric, a mere literary exercise divorced from political influence, became ornate and flowery in manner.

Late forms of poetry. The creative period of the Hellenistic Age was practically contained within the span of the 3rd century BC. To this period belonged three outstanding poets: Theocritus, Callimachus, and Apollonius of Rhodes. Theocritus (*c.* 310–250), born at Syracuse, is best known as the inventor of bucolic mime, or pastoral poetry, in which he presented scenes from the lives of shepherds and goatherds in Sicily and southern Italy. He also dramatized scenes from middle-class life; and in his second idyll the character Simaetha, who tries by incantations to recover the love of the man who has deserted her, touches the fringe of tragedy. He also used another Hellenistic form, the epyllion, a short scene of heroic narrative poetry in which heroic stature is often reduced by playful realism and delicate psychology. In his hands the hexameter attained a lyric purity and sweetness unrivaled elsewhere. He was the first of the nature poets, succeeded by Moschus and Bion.

Callimachus (flourished about 260) was a scholar as well as a poet. His most famous work, of which substantial fragments survive, was the *Aitia,* an elegiac poem describing the origins of various rites and customs. It was heavy with learning but diversified by passages of entertaining narrative. His six hymns show immense poetic expertise but no religious feeling, for the gods of Olympus had long since become obsolete. Callimachus also wrote epigrams, and fragments survive of *iambi* ("iambs"). The form was widely used throughout the 3rd century to denounce the vanities of the world. Sometimes, in a mixture of prose and verse, these pieces had links with satire; and their chief exponents were Bion the Borysthenite, Menippus of Gadara, Cercidas of Megalopolis, and Phoenix of Colophon.

The *Aitia* of Callimachus

Callimachus avoided epic in favour of the greater intensity possible in shorter works. The last surviving classical Greek epic was written by his successor at Alexandria, Apollonius of Rhodes (born about 295). Apollonius' account of the voyage of the Argonauts is so full of local legend that the coherence of the poem is lost; but the story of Medea's wild passion for Jason, the leader of the Argonauts, is marked by a new sort of romantic awareness that is fully realized in the episode of Dido's passion for Aeneas in Virgil's *Aeneid.*

The desire to combine learning with poetry led to the revival of didactic verse. The *Phaenomena* of Aratus of Soli (*c.* 315–*c.* 245) is a versification of a treatise on the stars by Eudoxus of Cnidus (*c.* 390–*c.* 340). Chance has preserved the poems of Nicander (probably 2nd century) on the unlikely subjects of cures for bites and antidotes to poisons.

The mimes of Herodas (3rd century), short realistic sketches of low life in iambic verse, have affinities with the non-pastoral mimes of Theocritus. They perhaps give a hint as to the character of the literature of popular entertainment, now largely lost. Mime, especially pantomime, was the main entertainment throughout the early Roman Empire.

After the middle of the 3rd century, poetic activity largely died away, though the great period of scholarship at Alexandria and at Pergamum was still to come. The names of a few poets are known: Euphorion (born about 275) of Chalcis and Parthenius (flourished 1st century BC), the teacher of Virgil. Thereafter Greek poetry practically ceased, apart from a sporadic revival in the 4th century AD. An exception exists in the case of epigrammatic poetry in elegiac couplets, surviving mainly in two compilations, the Planudean and Palatine anthologies.

Late forms of prose. Almost all of the great mass of Hellenistic prose—and later prose, historical, scholarly, and scientific—has perished. Among historians Polybius (*c.* 200–*c.* 118 BC), the most outstanding, has survived in a fragmentary condition. Present at Rome when it was succumbing to the first influences of Greek literature, he wrote mainly of events of which he had direct experience, often with great insight; his work covered the period from 264 to 146. Diodorus Siculus' universal history (1st century BC) is important for the sources quoted there. The most considerable of lost historians was Timaeus (*c.* 356–*c.* 260), whose history of the Greeks in the west down to 264 provided Polybius with his starting point. Later historians were Dionysius of Halicarnassus (flourished about 20 BC); Appian of Alexandria (2nd century AD), who wrote on Rome and its conquests; and Arrian (*c.* AD 96–*c.* 180) from Bithynia, who is the most valuable source on Alexander the Great.

The most important works of criticism, of which little has survived, were by Dionysius of Halicarnassus and the obscure Longinus. Longinus' treatise *On the Sublime* (*c.* AD 40) is exceptional in its penetrating analysis of creative literature. The *Bibliotheca* attributed to Apollodorus (*c.* 180 BC) is a handy compendium of mythology.

Scientific work such as the astronomy and geography of Eratosthenes (*c.* 276–*c.* 194) of Alexandria is known mainly from later summaries; but much that was written by the mathematicians, especially Euclid (flourished *c.* 300 BC) and Archimedes (*c.* 287–212), has been preserved.

Much survives of the writings of the physician Galen (AD 129–199). His contemporary Sextus Empiricus is an important source for the history of Greek philosophy. The survey of the Mediterranean by Strabo in the time of Augustus preserved much valuable information; and so, in a more limited field, did the description of Greece by Pausanias (2nd century AD). Greek achievement in astronomy and geography was summed up in the work of Ptolemy of Alexandria in the 2nd century AD.

Greek became the language of the large settlement of Jews at Alexandria, and the Septuagint, the Greek version of the Old Testament, was completed by about the end of the 2nd century BC. Much of the Apocrypha was composed in Greek, and the New Testament was written in popular Greek (Koine). Of the early Christian writers in Greek the most notable were Clement of Alexandria (*c.* AD 150–*c.* 215) and Origen (*c.* AD 185–*c.* 254), together with Clement of Rome and Ignatius of Antioch.

Plutarch's *Lives*

The *Parallel Lives* of famous Greeks and Romans by Plutarch (*c.* AD 46–*c.* 119) of Chaeronea in Boeotia was for centuries one of the formative books for educated Europeans. Great figures from an idealized past are presented for the edification of the lesser men of his own day; and the anecdotes with which the *Lives* abound are of various degrees of credibility. They belong to biography rather than to history, though they are an important source for historians. A number of shorter works on a wide variety of subjects have come down under the Latin title *Moralia* (Greek *Ethica*), which show the intellectual tide of Greece on the ebb.

There was much concern over a question that had been argued ever since the days when Athens had ceased to be a free city: to what extent was Attic prose a norm that writers and especially orators were bound to follow? Many had shunned it in favour of a more ornamental Asiatic style. But at the end of the 1st century AD there was a revival of the Attic dialect. Speeches and essays were written for wide circulation. This revival is known as the Second Sophistic movement, and chief among its writers were Dion Chrysostom (1st century AD), Aelius Aristides (2nd century), and Philostratus (early 3rd century). The only writer of consequence, however, was Lucian (*c.* 120–*c.* 190). His works are mainly slight and satirical; but

his gift of humour, even though repetitive, cannot be denied. *Lives and Opinions of Eminent Philosophers* was a valuable work of the 3rd century by Diogenes Laërtius, a writer otherwise unknown.

Philosophical activity in the early empire was mainly confined to moralizings based on Stoicism, a philosophy advocating a life in harmony with nature and indifference to pleasure and pain. Epictetus (born about AD 55) influenced especially the philosophic Roman emperor Marcus Aurelius (121–180), whose *Meditations* have taken their place beside works of Christian devotion. Many of Plutarch's *Moralia* were Platonic, with vaguely mystical tendencies; but Plotinus (*c.* 205–260/270) was the last major thinker in the classical world, giving new direction to Platonic and Pythagorean mysticism.

Erotic romances

The latest creation of the Greek genius was the novel, or erotic romance. It may have originated as early as the 1st century BC; but its roots reach back to such plays of triumphant love as the lost *Andromeda* of Euripides, to the New Comedy, to Xenophon's daydreams about the education of Cyrus, and to the largely fictitious narratives that were one extreme of what passed for history from the 3rd century BC onward. Of these last, the best known examples are the Alexander romances, a wildly distorted and embroidered version of the exploits of Alexander the Great, which supplied some of the favourite reading of the Middle Ages. Erotic elegy and epigram may have contributed something and so may the lost *Milesian Tales* of Aristides of Miletus (*c.* 100 BC), though these last appear to have depended on a pornographic interest that is almost completely absent from the Greek romances. Only fragments survive of the Ninus romance (dealing with the love of Ninus, legendary founder of Nineveh), which was probably of the 1st century BC; but full-length works survive by Chariton (2nd century AD), Heliodorus (3rd century AD), Xenophon (2nd or 3rd century AD) of Ephesus, and Achilles Tatius (2nd century AD). All deal with true lovers separated by innumerable obstacles of human wickedness and natural catastrophe and then finally united. *Daphnis and Chloe* by Longus (between 2nd and 3rd century AD) stands apart from the others because of its pastoral, rather than quasi-historical, setting. The works of Dictys Cretensis and Dares Phrygius belong to the same period. They claim to give a pre-Homeric account of the Trojan War. The Greek originals are almost wholly lost, but the Latin version was for the Middle Ages the main source for the story of Troy. (D.W.L./Ed.)

Byzantine literature

GENERAL CHARACTERISTICS

Byzantine literature may be broadly defined as the Greek literature of the Middle Ages, whether written in the territory of the Byzantine Empire or outside its borders. By late antiquity many of the classical Greek genres, such as drama and choral lyric poetry, had long been obsolete, and all Greek literature affected to some degree an archaizing language and style, perpetuated by a long-established system of education in which rhetoric was a leading subject. The Greek Church Fathers were the products of this education and shared the literary values of their pagan contemporaries. Consequently the vast and imposing Christian literature of the 3rd to 6th centuries, which established a synthesis of Hellenic and Christian thought, was largely written in a language already far removed from that spoken by all classes in everyday life, and indeed from that of the New Testament. This diglossy—the use of two very different forms of the same language for different purposes—marked Byzantine culture for 1,000 years; but the relations between the high and low forms changed with the centuries. The prestige of the classicizing literary language remained undiminished until the end of the 6th century; only some popular saints' lives and world chronicles escaped its influence. In the ensuing two and a half centuries, when the very existence of the Byzantine Empire was threatened, city life and education declined, and with them the use of classicizing language and style. With the political recovery of the 9th and 10th centuries began a literary revival, in which a conscious attempt was made to recreate the Hellenic-Christian culture of late antiquity. Simple or popular language was despised; many of the early saints' lives were rewritten in inflated and archaizing language and style. By the 12th century the cultural self-assurance of the Byzantines enabled them to develop new literary genres, including romantic fiction, in which adventure and love are the main motifs, and satire, which occasionally made use of imitations of spoken Greek. The period from the Fourth Crusade (1204) to the capture of Constantinople by the Ottoman Turks (1453) saw both a vigorous revival of narrowly imitative, classicizing literature, as the Byzantines sought to assert their cultural superiority over the militarily and economically more powerful West, and at the same time the beginning of a flourishing literature in an approximation to vernacular Greek. But this vernacular literature was limited to poetic romances, popular devotional writing, and the like. All serious writing continued to make use of the prestigious archaizing language of learned tradition.

Relation to classical and Christian literatures

Byzantine literature's two sources, classical and Christian, each provided a series of models and references for the Byzantine writer and reader. Often both were referred to side by side: for example, the emperor Alexius Comnenus defended his seizure of church property to pay his soldiers by referring to the precedents of Pericles and the biblical king David. Much of Byzantine literature was didactic in tone, and often in content too. And much of it was written for a limited group of educated readers, who could be counted upon to understand every classical or biblical allusion and to appreciate every figure of rhetoric. Some Byzantine genres would not be considered of literary interest today, but instead seem to belong to the domain of technical writing. This is true in particular of the voluminous writings of the Church Fathers, such as Athanasius, Gregory of Nazianzus, Basil, John Chrysostom, Cyril of Alexandria, and Maximus the Confessor.

PRINCIPAL FORMS OF WRITING

Nonliturgical poetry. Poetry continued to be written in classical metres and style. But the sense of appropriateness of form to content was lost. An example is the transitional work of Nonnus, a 5th-century Egyptian-born Greek who eventually converted to Christianity. His long poem *Dionysiaca* was composed in Homeric language and metre, but it reads as an extended panegyric on Dionysus rather than as an epic. Nonnus is plausibly credited with a paraphrase, in similar metre and style, of the Gospel According to St. John, thereby fusing classical and Christian traditions. Several short narrative poems in Homeric verse, of mythological content, were composed by contemporaries of Nonnus. Paul the Silentiary in the mid-6th century used the same Homeric form for a long descriptive poem on the Church of the Divine Wisdom (Hagia Sophia) in Constantinople. Many brief occasional poems were written in hexameters or elegiac couplets until the late 6th century. But changes in the phonology of Greek, and perhaps declining educational standards, made these metres difficult to handle. A cleric, George the Pisidian, wrote long narrative poems on the wars of the emperor Heraclius (610–641), as well as a poem on the six days of the creation, in iambic trimeters (12-syllable lines, consisting in principle of three pairs of iambic feet, each of a short syllable followed by a long). His example was followed by Theodosius the Deacon in his epic on the recapture of Crete from the Arabs in the 10th century. This 12-syllable line became the all-purpose metre in the middle and later Byzantine periods and was the vehicle for narrative, epigram, romance, satire, and moral and religious edification. From the 11th century it found a rival in a 15-syllable stressed line, which was used by the monk Symeon the New Theologian in many of his mystical hymns and which became a vehicle for court poetry in the 12th century. It was also used by the metropolitan Constantine Manasses for his world chronicle and by the anonymous redactor of the epic romance of *Digenis Akritas*. It was in this metre, which followed no classical models, that the early vernacular poems were written, such as the romances of *Callimachus and Chrysorrhoe*, *Belthandros and Chrysantza*, the Byzantine *Achilleid* (the

Developments in Byzantine poetry

hero of which has nothing in common with Homer's Achilles but his name), and the *Romance of Belisarius.* These are the most significant works of genuine fiction in Byzantine literature. Many of these poems were adaptations or imitations of medieval Western models: examples are *Phlorios and Platziaphlora* (the Old French *Floire et Blancheflor*), *Imberios and Margarona,* and *Apollonius of Tyre,* each a romantic narrative. The epic genre is represented by a long unpublished poem on the Trojan War, adapted from the *Roman de Troie* of the 12th-century French poet Benoît de Sainte-Maure. This openness to the Latin West was new. But even when they were based on Western models, Byzantine poems differed in tone and expression from their exemplars. Most of this vernacular poetry cannot be dated more precisely than to the 13th or 14th century.

Much Byzantine poetry is rather unimaginative, long-winded, and tedious. But some poets show a genuine vein of inspiration, for instance, John Geometres (10th century) or John Mauropous (11th century), or remarkable technical brilliance, such as Theodore Prodromus (12th century), or Manuel Philes (14th century). The ability to write passable verse was widespread in literate Byzantine society, and poetry—or versification—was greatly appreciated.

Liturgical poetry. From the earliest times song—and short rhythmic stanzas (*troparia*) in particular—had formed part of the liturgy of the church. Poems in classical metre and style were composed by Christian writers from Clement of Alexandria and Gregory of Nazianzus to Sophronius of Jerusalem. But the pagan associations of the genre, as well as the difficulties of the metre, made them unacceptable for general liturgical use. In the 6th century elaborate rhythmical poems (*kontakia*) replaced the simpler *troparia.* They owed much to Syriac liturgical poetry. In form the *kontakion* was a series of up to 22 rhythmical stanzas, all constructed on the same accentual pattern and ending with the same short refrain. In content it was a narrative homily on an event of biblical history or an episode in the life of a saint. There was often a marked dramatic element. Rich in imagery, complex in structure, and infinitely variable in rhythm, the new liturgical poetry can be compared with the choral lyric of ancient Greece. The greatest composer of *kontakia* was Romanos Melodos (Romanos the Melode; early 6th century), a Syrian probably of Jewish origin. In the late 7th century the *kontakion* was replaced by a longer liturgical poem, the *kanōn,* consisting of eight or nine odes, each of many stanzas and each having a different rhythmic and melodic form. The *kanōn* was a hymn of praise rather than a homily. Its great length encouraged repetition and inflation, and a more ornamental style of singing enhanced the importance of the music at the expense of the words. The most noteworthy composers of *kanones* were Andrew of Crete, John of Damascus, Theodore Studites, Joseph the Hymnographer, and John Mauropous. No new hymns were added to the liturgy after the 11th century, but *kanones* continued to be composed as a literary exercise. The original music of *kontakia* and *kanones* alike is lost.

The *kontakion*

The *kanōn*

Historical works. Conscious as they were of their classical and biblical past, the Byzantines wrote much history. Until the early 7th century a series of historians recounted the events of their own time in classicizing style, with fictitious speeches and set descriptive pieces, in a genre that owed much to the classical Greek historians Thucydides and Polybius. Procopius, Agathias, Peter the Patrician, Menander Protector, and Theophylactus Simocattes each took up where a predecessor left off. Thereafter this vein virtually ran dry for 250 years. The revival of cultural confidence and political power in the late 9th century saw a revival of classicizing history, with an interest in human character—Plutarch was often the model—and the causes of events. Joseph Genesius in the 10th century and the group of historical writers known collectively as the Continuators of Theophanes recorded, not without partiality, the origin and early days of the Macedonian dynasty. From then until the later 14th century there was never a generation without its historian. The most noteworthy historians were Symeon the Logothete and Leo the Deacon in the 10th century; Michael Psellus, Michael Attaleiates, and John Scylitzes in the 11th century; Anna Comnena, John Cinnamus, and Nicetas Choniates in the 12th century; George Acropolites and George Pachymeres in the 13th century; and Nicephorus Gregoras and the emperor John Cantacuzenus in the 14th century. The last days of the Byzantine Empire were recounted from very different points of view by George Sphrantzes, the writer known simply as Ducas (who was a member of the former Byzantine imperial house of that name), Laonicus Chalcocondyles, and Michael Critobulus in the second half of the 15th century.

Another kind of interest in the past was satisfied by world chronicles, beginning with the creation or some early biblical event. Often naively theological in their explanation of causes, black-and-white in their depiction of character, and popular in language, they helped the ordinary Byzantine to locate himself in a scheme of world history that was also a history of salvation. The *Chronographia* of John Malalas in the 6th century and the Paschal Chronicle (*Chronicon Paschale*) in the 7th century were succeeded by those of Patriarch Nicephorus at the end of the 8th century, Theophanes the Confessor in the early 9th century, and George the Monk in the late 9th century. Such chronicles continued to be written in later centuries, sometimes with critical and literary pretensions, as in the case of John Zonaras, or in vaguely romanticized form in verse, as in the case of Constantine Manasses.

World chronicles

The importance that Byzantine rulers attached to history is attested by the vast historical encyclopaedia compiled on the orders of Constantine VII (913–959) in 53 volumes, of which only meagre fragments remain.

Rhetoric. Though there was no opportunity for political or forensic oratory in the Byzantine world, the taste for rhetoric and the appreciation of well-structured language, choice figures of speech and thought, and skillful delivery remained undiminished in Byzantine society. From the 10th century onward survives a vast body of encomiums, funeral orations, memorial speeches, inaugural lectures, addresses of welcome, celebrations of victory, and miscellaneous panegyrics. This outpouring of polished rhetoric played an important role in the formation and control of public opinion in the limited circles where opinions mattered and occasionally served as a vehicle of genuine political controversy. To this same domain belong the myriad Byzantine letters, often collected and edited by their author or a friend. These letters were not intended to be either private or informative—real information was conveyed orally by the bearer—but they were important in maintaining networks of contact among the elite as well as in providing refined aesthetic pleasure. (R.B.)

Modern Greek literature (after 1453)

POST-BYZANTINE PERIOD

After the Turkish capture of Constantinople in 1453, Greek literary activity continued almost exclusively in those areas of the Greek world under Venetian rule. Thus Cyprus, until its capture by the Turks in 1571, produced a body of literature in the local dialect, including the 15th-century prose chronicle *Recital Concerning the Sweet Land of Cyprus* by Leóntios Machairás and a collection of translations and imitations in elaborate verse forms of Italian poems by Petrarch and others. Crete, which remained in Venetian hands until 1669, became the centre of the greatest flowering of Greek literature between the fall of Constantinople and the foundation of the modern Greek state. There a number of authors developed the Cretan dialect into a rich and subtle medium of expression. In it were written a number of tragedies and comedies, a single pastoral tragicomedy, and a single, anonymous religious drama, *The Sacrifice of Abraham,* mostly based on Italian models. The leading playwright was Geórgios Chortátsis. In the first half of the 17th century Vitséntsos Kornáros composed his romance, *Erotókritos.* These Cretan authors composed their works almost entirely in the 15-syllable iambic verse of the Greek folk song, whose modes of expression influenced them deeply.

In the Ottoman-ruled areas of Greece the folk song, which concisely and unsentimentally conveyed the aspira-

tions of the Greek people of the time, became practically the sole form of literary expression.

Toward the end of the 18th century, however, a number of intellectuals emerged who, under the influence of European ideas, set about raising the level of Greek education and culture and laying the foundations of an independence movement. The participants in this "Greek Enlightenment" also brought to the fore the language problem, each promoting a different form of the Greek language for use in education. The leading Greek intellectual of the early 19th century was the classical scholar Adamántios Koraïs, who in voluminous writings on Greek language and education, argued for a form of Modern Greek "corrected" according to the ancient rules.

The Greek language problem

INDEPENDENCE AND AFTER

Old Athenian School. The Greek state established as a result of the Greek War of Independence (1821–29) consisted only of a small section of the present-day Greek mainland and a few islands. Athens, which became the capital of Greece in 1834, soon came to be the chief cultural centre, gathering together writers from various areas, particularly Constantinople. The Soútsos brothers, Aléxandros and Panayótis, introduced the novel into Greece, but they are best known for their Romantic poetry, which as time went by moved gradually away from the Demotic ("popular"), or commonly spoken, language toward the Katharevusa ("purist") form institutionalized by Koraïs. The work of these writers, which relied greatly on French models, looks back to the War of Independence and the glorious ancient past. Their melancholy sentimentality was not shared by Aléxandros Rízos Rangavís, a verbose but versatile and not inconsiderable craftsman of Katharevusa in lyric and narrative poetry, drama, and the novel. By the 1860s and '70s, however, Athenian poetry was generally of poor quality and was dominated by a sense of despair and longing for death. Prose throughout the period was monopolized by the historical novel. Emmanuel Roídis' novel called *I Pápissa Ioánna* (1866; *Pope Joan*) is a hilarious satire on medieval and modern religious practices as well as a pastiche of the historical novel. Pávlos Kalligás, in *Thános Vlékas* (1855), treated contemporary problems such as brigandage. In *Loukís Láras* (1879; Eng. trans., *Loukis Laras*) Dimítrios Vikélas presented a less heroic view of the War of Independence.

Heptanesian School. Meanwhile more interesting developments had been taking place in the Ionian Islands (Heptanesos). During the 1820s two poets from the island of Zacynthus made their name with patriotic poems celebrating the War of Independence. One of these, Andréas Kálvos, who composed his odes in neoclassical form and archaic language, never wrote poetry afterward, while the other, Dhionísios Solomós, went on to become one of the greatest of modern Greek poets. Dealing with the themes of liberty, love, and death, Solomós embodied a profoundly Romantic sensibility in extraordinary fragments of lyrical intensity, which gave a new prestige to the Demotic language. Solomós' followers continued to cultivate the Demotic, particularly Antónios Mátesis, whose historical social drama, *O vasilikós* (1859; "The Basil Plant"), was the first prose work of any length to be written in the Demotic. Aristotélis Valaorítis continued the Heptanesian tradition with long patriotic poems inspired by the Greek national struggles.

Demoticism and folklorism, 1880–1922. From the 1880s onward the New Athenian School, inspired by the revived interest in folklore as a survival of ancient Greek culture, began to react against the sterile bombast of the Katharevusa versifiers, producing instead a more intimate poetry based on the language, customs, and beliefs of the Greek peasantry, and in particular on Greek folk songs.

The leading ideologist of this "demoticist" movement, which aimed to promote traditional popular culture at the expense of the pseudo-archaic pedantry fashionable in Athens, was Yánnis Psicháris (Jean Psichari), whose book *My Journey* (1888) was partly a fictionalized account of a journey around the Greek world and partly a belligerent manifesto arguing that the Demotic language should be officially adopted as a matter of national urgency. The demoticist movement inspired poets to enrich the Greek popular tradition with influences from abroad. Chief among these was Kostís Palamás, who dominated the literary scene for several decades with a large output of essays and articles and whose best poetry appeared between 1900 and 1910. In his lyric and epic poems he attempted to synthesize ancient Greek history and mythology with the Byzantine Christian tradition and modern Greek folklore in order to demonstrate the essential unity of Greek culture. Angelos Sikelianós continued this enterprise in effusive and powerful lyric poetry of a profoundly mystical nature.

In prose, the folklore cult fostered development of the short story, written initially in Katharevusa, with Demotic gradually taking over in the 1890s. These stories, and the novels that accompanied them, depicted scenes of traditional rural life, sometimes idealized and sometimes viewed critically by their authors. The pioneer of the Greek short story, Geórgios Vizyenós, combined autobiography with an effective use of psychological analysis and suspense. The most famous and prolific short-story writer, Aléxandros Papadiamándis, produced a wealth of evocations of his native island of Skiáthos imbued with a profound sense of Christian tradition and a compassion for country folk; his novel *I fónissa* (1903; *The Murderess*) is a fine study in psychological abnormality. The novel *O zitiános* (1896; *The Beggar*), by Andréas Karkavítsas, satirically depicts the economic and cultural deprivation of the rural population. From about 1910 this critical attitude is further reflected in the prose writing of Konstantínos Chatzópoulos and Konstantínos Theotókis. Meanwhile Grigórios Xenópoulos wrote novels with an urban setting and devoted considerable effort to drama, a medium that received a substantial boost from the demoticist movement.

The short story

One major figure defies categorization for it was outside Greece, in Alexandria, that Constantine Cavafy lived and wrote. His finely wrought, epigrammatic poems, with their tragically ironic views of Hellenistic and Byzantine history, contain daring, sensuous glimpses of homosexual love.

Literature from 1922. The Asia Minor Disaster of 1922, in which Greece's expansionist designs on the Ottoman Empire were finally thwarted, brought about a radical change in the orientation of Greek literature. Before committing suicide, Kóstas Kariotákis wrote some bitterly sarcastic poetry conveying the gap between the old ideals and the new reality.

The reaction against the defeatism of 1922 came with the Generation of 1930, a group of writers who began publishing around that date. They reinvigorated Greek literature by discarding the old verse forms in poetry and by producing ambitious novels that were intended to embody the spirit of the times. Both poets and novelists sought to combine European influences with the best of what was Greek. The restrained poetry of George Seféris skillfully married references to ancient mythology with pensive meditation on man's modern situation, while his finely written essays recast the Greek tradition according to his own priorities. Odysseus Elýtis celebrated the Aegean scenery as an ideal world of sensual enjoyment and moral purity. Each of these poets won the Nobel Prize for Literature, Seféris in 1963 and Elýtis in 1979. Yánnis Rítsos adopted various new modes of writing in his celebration of the Greek partisans in World War II, in long dramatic monologues spoken by characters from Greek mythology, and in laconic poems depicting everyday, but often ironically presented, scenes.

The Generation of 1930 produced some remarkable novels, among them Strátis Myrivílis' *I zoí en tafo* (1930; *Life in the Tomb*), a journal of life in the trenches in World War I; *Argo* (2 vol., 1933 and 1936) by Yórgas Theotokás, about a group of students attempting to find their way through life in the turbulent 1920s; and *Eroica* (1937) by Kosmás Políti, about the first encounter of a group of well-to-do schoolboys with love and death.

After World War II prose writing was dominated by novels reflecting the experiences of the Greeks during eight years of war (1941–49). Iánnis Berátis recounted his experiences of 1941 in an unemotional manner in *To Platy*

World War II and afterward

Potami (1946; "The Broad River"). In a trilogy of novels entitled *Akyvérnites polities* (1960–65; *Drifting Cities*), Stratís Tsírkas masterfully recreated the atmosphere of the Middle East in World War II. In the short story, Dimítris Chatzís painted ironic portraits of real and fictional characters in his native Ioánnina in the period before and during World War II, exposing their self-interested machinations.

Kazantzákis' work

Nevertheless, the most famous novelist of the period, the Cretan Níkos Kazantzákis, was a survivor from an earlier generation. In a series of novels, beginning with *Vios ke politía tou Aléxi Zorbá* (1946; *Zorba the Greek*) and continuing with his masterpiece *O Christos xanastavronete* (1954; *Christ Recrucified*), he embodied a synthesis of ideas from various philosophies and religions in larger-than-life characters who wrestle with great problems, such as the existence of God and the purpose of human life. Kazantzákis had earlier published his 33,333-line *Odísia* (1938; *Odyssey*), an epic poem taking up the story of Odysseus where Homer had left off. Pandelís Prevelákis published a number of philosophical novels set in his native Crete, the most successful being *O ílios tou thanátou* (1959; *The Sun of Death*), which shows a boy learning to come to terms with death.

During the 1960s Greek prose writers attempted to explore the historical factors underlying the contemporary social and political situation. In the novel *To tríto stefáni* (1962; *The Third Wedding*) by Kóstas Tachtsís, the female narrator tells the story of her life with venomous verve, unwittingly exposing the oppressive nature of the Greek family. Yórgos Ioánnou's part-fictional, part-autobiographical short prose pieces present a vivid picture of life in Thessaloníki (Salonika) and Athens from the 1930s to the 1980s.

No individual poets of the postwar generations tower above the rest; but Tákis Sinópoulos, Míltos Sachtoúris, and Manólis Anagnostákis, all marked by their wartime experiences of the 1940s, are among those with the greatest reputations. (P.A.M.)

BIBLIOGRAPHY

Ancient Greek literature: Among many surveys of Greek literature, the best are P.E. EASTERLING and B.M.W. KNOX, *The Cambridge History of Classical Literature,* vol. 1, *Greek Literature* (1985); JACQUELINE DE ROMILLY, *A Short History of Greek Literature* (1985; originally published in French, 1980), with an excellent bibliography; K.J. DOVER (ed.), *Ancient Greek Literature* (1980); and ALBIN LESKY, *A History of Greek Literature* (1966; originally published in German, 2nd ed., 1963). Still valuable are GILBERT MURRAY, *A History of Ancient Greek Literature* (1897, reprinted 1966); and MOSES HADAS, *A History of Greek Literature* (1950, reissued 1962). N.G.L. HAMMOND and H.H. SCULLARD (eds.), *The Oxford Classical Dictionary,* 2nd ed. (1970, reprinted 1977); and PAUL HARVEY (comp.), *The Oxford Companion to Classical Literature* (1937, reprinted with corrections 1980), are excellent reference resources. Topical studies include ROBERT GRAVES, *The Greek Myths,* 2 vol. (1955, reprinted with amendments, 2 vol. in 1, 1957); J.W.H. ATKINS, *Literary Criticism in Antiquity: A Sketch of Its Development,* 2 vol. (1934, reissued 1961); C.M. BOWRA, *Greek Lyric Poetry from Alcman to Simonides* (1961, reprinted 1967): MARGARETE BIEBER, *The History of the Greek and Roman Theater,* 2nd rev. ed. (1961); BRIAN VICKERS, *Comparative Tragedy,* vol. 1, *Towards Greek Tragedy: Drama, Myth, Society* (1973, reprinted 1979); R.R. BOLGAR, *The Classical Heritage and Its Beneficiaries* (1954, reprinted 1977); GILBERT HIGHET, *The Classical Tradition: Greek and Roman Influences on Western Literature* (1949, reissued 1985); and M.J. ANDERSON (ed.), *Classical Drama and Its Influences* (1965). Trends in scholarship are discussed in MAURICE PLATNAUER, *Fifty Years (and Twelve) of Classical Scholarship,* 2nd ed. (1968).

(D.W.L./Ed.)

Byzantine literature: The standard reference works are HANS-GEORG BECK, *Kirche und theologische Literatur im byzantinischen Reich,* 2nd ed. (1977), and *Geschichte der byzantinischen Volksliteratur* (1971); and HERBERT HUNGER, *Die hochsprachliche profane Literatur der Byzantiner,* 2 vol. (1978). The standard work on Byzantine liturgical poetry is EGON WELLESZ, *A History of Byzantine Music and Hymnography,* 2nd ed. rev. and enlarged (1961, reprinted 1971). GEORGE A. KENNEDY, *Greek Rhetoric Under Christian Emperors* (1983), is concerned with social aspects, particularly with the rhetorical method in education. A collection of studies on the vernacular poetry of the period is E.M. JEFFREYS and M.J. JEFFREYS, *Popular Literature in Late Byzantium* (1983). ALEXANDER KAZHDAN, *Studies on Byzantine Literature of the Eleventh and Twelfth Centuries* (1984), brings methods of Western medievalists to the study of Byzantine literature. Influences are traced in KENNETH M. SETTON, *The Byzantine Background to the Italian Renaissance* (1956), a reliable survey. Also useful is *Ihor Ševčenko, Ideology, Letters and Culture in the Byzantine World* (1982).

(R.B.)

Modern Greek literature: C.TH. DIMARAS (K.TH. DEEMARAS), *A History of Modern Greek Literature* (1972; originally published in Greek, 2 vol., 1948–49), is a comprehensive study that is especially useful on the history of ideas. LINOS POLITIS, *A History of Modern Greek Literature* (1973), is the best general survey. Other useful works include CONSTANTINE A. TRYPANIS, *Greek Poetry: From Homer to Seferis* (1981), covering the entire Greek poetic tradition; RODERICK BEATON, *Folk Poetry of Modern Greece* (1980); PHILIP SHERRARD, *The Marble Threshing Floor: Studies in Modern Greek Poetry* (1956, reprinted 1981); EDMUND KEELEY and PETER BIEN (eds.), *Modern Greek Writers: Solomos, Calvos, Matesis, Palamas, Cavafy, Kazantzakis, Seferis, Elytis* (1972); EDMUND KEELEY, *Modern Greek Poetry: Voice and Myth* (1983); and PETER BIEN, *Three Generations of Greek Writers: Introductions to Cavafy, Kazantzakis, Ritsos* (1983). See also MARY P. GIANOS (comp.), *Introduction to Modern Greek Literature: An Anthology of Fiction, Drama, and Poetry* (1969).

(P.A.M.)

Biological Growth and Development

Although the essence of growth is an increase in the size or the amount of an entity, it is useful to distinguish between growth of living things and that of inanimate objects. The growth of inanimate objects, such as the process that occurs during the formation of crystals, is limited in an important way. The crystals can increase in size but are unable to reproduce themselves; living things, on the other hand, not only increase in size but also reproduce themselves. Growth of living things always occurs by an increase in the number or size of the basic units of organisms, cells, and always eventually includes reproduction. To appreciate fully the way in which living things grow and the ways this growth is regulated requires a consideration of the way in which cells increase in size and number.

The word growth is often used synonymously with development, and, indeed, increase in size is a striking feature of development. But development, in this context, may also include changes in the types of cell specialization (differentiation) and extensive movements of cells. Movements of cells and tissues during embryonic development are partially responsible for the form of the body and various organs.

This article is concerned with the processes of growth and development of living organisms from the moment that fertilization takes place through the processes of maturation and senescence. (Death in the biological sense and its other aspects are treated in detail in the article DEATH.)

This article is divided into the following sections and subsections:

Growth 362
- The process of growth 362
  - Types of growth
  - Normal and abnormal growth
- Factors that regulate growth 363
  - Environmental factors
  - Internal factors
- The dynamics of growth 364
  - Measurement of growth
  - The study of growth
- Aberrations of growth: biological malformation 365
  - Plant malformations
  - Animal malformations
- Biological regeneration 367
  - Modes of regeneration
  - The regeneration process
  - The range of regenerative capability

General features of biological development 371
- The scope of development 372
  - Types of development
  - General systems of development
  - Constituent processes of development
- Control and integration of development 376
  - Phenomenological aspects
  - Analytical aspects
- Development and evolution 377
  - Effect on life histories
  - Genetic assimilation

Plant development 379
- General features 379
  - Life cycles
  - Body plans
  - Preparatory events
- Early development: from zygote to seedling 381
  - Embryo formation
  - Germination and early growth
- Later development: the sporophyte plant body 384
  - Continuation of organ formation
  - The shoot system and its derivatives
  - The root system and its derivatives
- Correlations in plant development 388
  - Coordination of shoot and root development
  - Determination of mature form
  - Seasonal adaptations

Animal development 390
- General features 390
  - Reproduction and development
  - Preparatory events
- Early development 391
  - Embryo formation
  - Embryonic adaptations
- Organ formation 396
  - Primary organ rudiments
  - Organogenesis and histogenesis
- Ectodermal derivatives 398
  - The nervous system
  - The epidermis and its outgrowths
- Mesodermal derivatives 400
  - The body muscles and axial skeleton
  - The appendages: tail and limbs
  - Excretory organs
  - Circulatory organs
  - Reproductive organs
- Endodermal derivatives 404
  - The alimentary canal
  - The pharynx and its outgrowths
  - The liver, pancreas, and lungs
- Postembryonic development 404
  - The larval phase and metamorphosis
  - Direct development
- Maturity and death 407

Human growth and development 407
- Human embryology: early stages 407
  - From fertilization to placentation
  - Formation of the three primary germ layers
  - Growth and differentiation
  - Embryonic acquisition of external form
- Abnormal development 413
  - Multiple births
  - Fetal deviations
  - Teratology
- Development of organs 414
  - Ectodermal derivatives
  - Mesodermal derivatives
  - Endodermal derivatives
- Postnatal development 418
  - Types and rates of human growth
  - Types of growth data
- Development at puberty 421
  - Alterations in growth rate
  - Normal variations
- Hormones and growth 424

Aging and senescence 424
- Life-span 424
  - Measurement of life-span
  - Plants
  - Animals
  - Human life-span
- Aging: general considerations 429
  - Biological theories of aging
  - Natural history of aging
  - Senescence in plants
  - Senescence in mammals
  - Aging at the molecular and cellular levels
  - Internal and external causes of aging
- Human aging 435
  - Effect of aging of the body systems
  - Psychological aspects of aging

Bibliography 439

GROWTH

The process of growth

The increases in cell size and number that take place during the life history of an organism are seldom random; rather, they occur according to a plan that eventually determines the size and shape of the individual. Growth may be restricted to special regions of the organism, such as the layers of cells that divide and increase in size near the tip of the plant shoot. Or the cells engaged in growth may be widely distributed throughout the body of the organism, as in the human embryo. In the latter case, the rates of cell division and of the increase in cell size differ in different parts. That the pattern of growth is predetermined and regular in plants and animals can be seen in the forms of adults. In some organisms, however, notably the slime molds (see PROTOPHYTES: *Slime molds*), no regular pattern of growth occurs, and a formless cytoplasmic mass is the result.

The rate of growth of various components of an organism may have important consequences in its ability to adapt to the environment and hence may play a role in evolution. For instance, an increase in the rate of growth of fleshy parts of the fish fin would provide an opportunity for the fish to adapt more easily to terrestrial locomotory life than could a fish without this modified fin. Without disproportionate growth of the fin—ultimately resulting from random changes in the genetic material (mutations)—the evolution of limbs through natural selection might have been impossible.

TYPES OF GROWTH

In cells. The increase in size and changes in shape of a developing organism depend on the increase in the number and size of cells that make up the individual. Increase in cell number occurs by a precise cellular reproductive mechanism called mitosis. During mitosis the chromosomes bearing the genetic material are reproduced in the nucleus, and then the doubled chromosomes are precisely distributed to the two daughter cells, one of each chromosomal type going to each daughter cell. Each end of the dividing cell receives a complete set of chromosomes before the ends separate. In animal cells this is a pinching off (cytokinesis) of the cell membrane; in plant cells a new cellulose wall forms between the new cells.

During the period of cell life preceding the actual distribution of chromosomes, the mother cell often grows to twice its original size. Hence, a cycle consisting of cell growth and cell division is established. Cell growth—an increase in cytoplasmic mass, chromosome number, and cell surface—is followed by cell division, in which the cytoplasmic mass and chromosomes are distributed to the daughter cells. An increase in cytoplasmic mass does not always occur during cell-division cycles, however. During the early development of an embryo, for example, the original egg cell, usually a very large cell, undergoes repeated series of cell divisions without any intervening growth periods; as a result, the original egg cell divides into thousands of small cells. Only after the embryo can obtain food from its environment does the usual pattern of growth and mitosis occur.

In plants. The fact that most plant cells undergo extensive size increase unaccompanied by cell division is an important distinction between growth in plants and in animals. Daughter cells arising from cell division behind the tip of the plant root or shoot may undergo great increases in volume. This is accomplished through uptake of water by the cells; the water is stored in a central cavity called a vacuole. The intake of water produces a pressure that, in combination with other factors, pushes on the cellulose walls of the plant cells and is responsible for the increase in length and girth of the cells and of the plant. In plants, much of the size increase occurs after cell division and results primarily from an increase in water content of the cells without much increase in dry weight.

ıe role
water in
ant-cell
owth

The very young developing plant embryo has many cells distributed throughout its mass that undergo the cycle of growth and cell division. As soon as the positions of the root tip, shoot tip, and embryonic leaves become established, however, the potential for cell division becomes restricted to cells in certain regions called meristems. One meristematic centre lies just below the surface of the growing root; all increases in the number of cells of the primary root occur at this point. Some of the daughter cells remain at the elongating tip and continue to divide. Other daughter cells, which are left behind in the root, undergo the increase in length that enables the new root to push deeper into the soil. The same general plan is evident in the growing shoot of higher plants, in which a restricted meristematic region at the tip is responsible for the formation of the cells of the leaves and stem; cell elongation occurs behind this meristematic centre. The young seedling secondarily develops cells associated with the vascular strands of phloem and xylem—tissues that carry water to the leaves from the soil and sugar from the leaves to the rest of the plant. These cells can divide again, providing new cell material for development of a woody covering and for more elaborate vascular strands. Hence, the growth of higher plants—*i.e.*, those aspects involving both the pattern of stems, leaves, and roots and the increase in bulk—results primarily from cell division at the meristem followed by a secondary increase in size because of water uptake. These activities occur throughout the period of plant growth.

In animals. The growth of animals is more restricted in time than is that of plants, but cell division is more generally distributed throughout the body of the organism. Although the rate of cell division differs in different regions, the capacity for cell division is widely distributed in the developing embryo. Increase in size is rapid during the embryonic period, continues at a reduced rate in juveniles, and thereafter is absent. Cell division and size increase continue, however, even after increase in total body size no longer occurs. Because these events are balanced by cell death, post-juvenile increase in cell number is primarily a replacement phenomenon. Height increase in mammals is limited by cessation of cell division and bone deposition in the long bones. The long juvenile period of growth in humans is unusual, most higher animals attaining mature size soon after the end of embryonic development. Some organ systems undergo little cell division and growth after birth; for instance, all of the germ cells (precursors of egg cells) of the female are formed by the time of birth. Similarly, all of the nerve cells of the brain are formed by the end of the embryonic period. Further increase in the size of the nervous system occurs by outgrowth of nerve fibres and deposition of a fatty insulation material along them. Although the greatest increase in size of nerve cells occurs, as in plant cells, after the cessation of cell division, the nerve fibre outgrowth in animals represents a true increase in the amount of cytoplasm and cell surface and not just an uptake of water.

Some organs retain the potential for growth and cell division throughout the life-span of the animal. The liver, for example, continues to form new cells to replace senescent and dying ones. Although cell division and growth occur throughout the liver, other organs have a special population of cells, called stem cells, that retain the capacity for cell division. The cells that produce the circulating red cells of mammalian blood are found only in the marrow of the long bones. They form a permanent population of dividing cells, replacing the red cells that continuously die and disappear from the circulation.

The rates of both growth and cell division can vary widely in different body parts. This differential increase in size is a prime factor in defining the shape of an organism.

NORMAL AND ABNORMAL GROWTH

Tumours. When growth is not properly regulated, anomalies and tumours may result. If the increase in the number of liver cells is abnormal, for example, tumours

Features associated with tumour growth

of the liver, or hepatomas, may result. In fact, one feature of malignant tumours, or cancers, is the absence of the usual growth patterns and rates. The cells of malignant tumours, in addition to having abnormal growth rates, have altered adhesive properties, which enable them to detach easily from the tumour; in this way the cells may spread to other parts of the body (metastasize) and grow in unusual locations. It is the growth of tumours in places other than the organ of origin that usually causes the death of an organism. Tumours may vary widely in their growth rates. They may grow very rapidly or so slowly that the rate approaches that of normal cell division in adult tissues. Tumours are not only characterized by an increase in the rate of cell division but also by abnormal patterns of growth. The new cells formed in the tumour are not organized and incorporated into the structure of the organ and may form large nodules. These abnormal growths may present no medical problems (*e.g.,* moles) or may cause disastrous effects, as is the case of the pressure on the brain caused by a tumorous mass of the meningeal covering of the brain. (For additional information about abnormal growth, see below *Aberrations of growth: biological information.*)

Regeneration. Not all abnormal growths are tumours. If a tree is partially burned, cells below the bark produce a new covering for the exposed vascular strands. Growth may not be normal, and an obvious scar or growth of the new bark is apparent. Similarly, if the skin of a mammal is severely injured, the repair, although abnormal and imperfect, causes the organism no physiological difficulty. Many organisms possess the ability to regrow, or regenerate, with varying degrees of perfection, parts of the body that are lost or injured. Salamanders possess remarkable powers of regeneration, being able to form new eyes or a new limb if the original is lost. Lizards can regenerate a new tail; even humans can regenerate parts of the liver. The reasons for the differences in regenerative powers in different animals remain a fascinating mystery of great practical importance. When regeneration does occur, some specialized cells usually lose their specialized characteristics and enter a period of an increased rate of cell division; subsequently, the new cells respecialize into the tissues of the original body part. Plants whose tops are lost as in pruning can also sometimes form new meristematic centres from dormant tissues and produce new shoots. (For additional information see below *Biological regeneration.*)

Compensatory growth. Many organs of animals occur in pairs, and if one is lost the remaining member increases in size, as if responding to the demands of increased use. If one of the two kidneys of a human is removed, for example, the other increases in size. This is called a compensatory reaction and may occur either by some increase in cell size (hypertrophy), by an increase in the rate of cell division (hyperplasia), or both. Although an increase in cell number is primarily responsible for the compensatory reaction of the kidney, the number of individual filtration units (glomeruli) does not increase. Hence, cell division increases the size of glomeruli but not the total number. Some of the most striking examples of increases in cell size in animals take place during stimulation of endocrine organs, which secrete regulatory substances called hormones; when the thyroid gland is stimulated, for example, the individual cells of the gland may increase dramatically in size.

Compensatory growth in the kidney

Factors that regulate growth

ENVIRONMENTAL FACTORS

Temperature. The environment in which an organism lives plays an important role in modifying the rate and extent of growth. Environmental factors may be either physical (*e.g.,* temperature, radiant energy, and atmospheric pressure) or chemical. Organisms and the cells of which they are composed are extremely sensitive to temperature changes; as the temperature decreases, the biochemical reactions necessary for life occur more slowly. A lowering of the temperature by 10° C (18° F) slows metabolism at least twofold and often more.

The width of trees increases partly by cell division and enlargement of secondary meristematic tissue below the bark. During the cold of winter, cell division and enlargement may cease completely; but during the spring renewed growth occurs. This intermittent growth is influenced by temperature, light, and water. The amount of growth may decrease considerably if the spring is cold, if day length is changed by obstructions blocking the sunlight, or if a drought occurs. In fact, the width of the growth rings visible on the surface of the cut tree trunk provides a partial history of climatic conditions, the spacing of the growth rings of different size having been correlated with known periods of drought and cold to provide reliable archaeological dating of various structures, as in the timbers used in Indian pueblos in the southwestern United States.

Temperature also affects both warm- and cold-blooded animals. Many warm-blooded (*e.g.,* bears) and cold-blooded (*e.g.,* frogs) vertebrates cease growing during the cold winter and simply enter an inactive or dormant state, which is characterized by a very low rate of metabolism. In animals that do not become dormant, increased demands for food consumption occur during cold periods to provide energy to maintain body temperature; this utilization of food energy may limit the energy available for size increase if food is in short supply.

Dormancy

Pressure. Because atmospheric pressure is relatively constant except in the mountains, it probably is of little importance in growth regulation. Increases in pressure in the ocean's depths may be significant, however, since it is known that increases in hydrostatic pressure interfere with cell division. Tissues of deep-sea fishes must have become adapted to such pressure effects, which have been little studied thus far. Movements of the terrestrial atmosphere—winds—may affect growth patterns in trees and shrubs, as is evident in the exotic shapes of certain conifers that grow along coastlines exposed to strong prevailing winds.

Light. Of all the physical factors, light plays the best understood and most dramatic role. Many of the effects of light on plant growth are obvious and direct. Light energy is the driving force for photosynthesis, the series of chemical reactions in green plants in which carbon dioxide and water form carbohydrates and upon which all life ultimately depends. Insufficient light causes death or retardation of growth in green plants. But light also has indirect effects of great importance. Green plants possess small amounts of a pigment called phytochrome that can exist in two forms. One form absorbs red light (660 millimicrons, or mμ; 1 mμ = 3.937×10^{-8} inch). When plants containing this pigment absorb red light, the pigment is converted to another form, which absorbs far-red light (730 mμ); the latter form can be converted back again to the original red absorbing form. These conversions have dramatic consequences; for example, red light inhibits stem elongation and lateral root formation but stimulates leaf expansion, chloroplast development, red flower coloration, and spore germination. Cycles of red and far-red light also can affect flower formation.

Indirect effects of light on plants

The effects of light on animals, although less obvious, may be important, as, for example, the effect of light on growth of the reproductive system of some animals. Increase in day length, hence in the amount of light, seems to initiate growth and development of the sex organs (gonads) in some birds during the spring. Curiously, the eyes are not the receptors for the light signal that activates the endocrine system to initiate growth of gonads; rather, cells deep in the brain are sensitive to the small amounts of light that pass directly through the thin skull of the bird.

Most animals show cyclic activity, or rhythms, in various important physical (*e.g.,* movement) and chemical (*e.g.,* respiration) events that are essential to the individual. These rhythms are often regulated by short exposure to light.

Chemical factors. Chemical factors of importance in the environment include the gases in the atmosphere and the water, mineral, and nutritional content of food. Plants require carbon dioxide, water, and sunlight for photosynthesis; drought slows plant growth and may even kill the plant. The effects of atmospheric contaminants—*e.g.,* oxides of nitrogen, hydrocarbons, and carbon monoxide—

are known to have deleterious effects on the growth and reproduction of both plants and animals.

Plants and animals require minerals and small amounts of elements such as zinc, magnesium, and boron. Nitrogen and phosphorus are provided to plants as nitrates and phosphates in the soil. Inadequate quantities of any nutritional factor in the soil result in poor plant growth and poor crop yields. Animals require oxygen, water, and elements from the environment. Because they are unable to synthesize sugars from carbon dioxide, animals must acquire these nutrients through the diet, either directly, by the consumption of plants, or indirectly, by the consumption of other animals that in turn have utilized plants as food. If the quality or quantity of this food is poor, either growth is retarded or death occurs (see NUTRITION).

Vitamins in diet

Vitamins, a class of compounds with a variety of chemical structures, are needed by animals in small amounts. Animals cannot synthesize all vitamins they require; those that cannot be synthesized must therefore be acquired in the diet, either from plants or from other animals that can synthesize the vitamin. Because certain vitamins are necessary in certain important metabolic reactions, vitamin deficiency during growth may have a variety of effects—stunting, malformation, disease, or death (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Vitamins*).

INTERNAL FACTORS

The organism is dependent on the environment for the raw materials for growth, but growth is also regulated internally. Because the size and form of plants and animals are under genetic control, events such as the rate and site of cell division and the extent of cell enlargement can be affected by mutations. It is not yet known, however, precisely how these factors, which are the ultimate determinants of growth, are controlled in individual cells.

One very important class of intrinsic growth regulators is that of the hormones. The principal plant hormone, auxin, is produced in the leaves; it moves by precise mechanisms, as yet poorly understood, to the other parts of the plant, controlling such processes as elongation of plant cells. Auxin somehow changes the characteristics of the rigid cell wall of the plant cell so that it becomes more flexible; the internal pressure within the cell then forces it to become larger. Other plant hormones may also play a role in the process; hormones such as cytokinins and gibberellins influence the rate of cell division in the meristems. Some dwarf plants can be stimulated to grow to normal size simply by applying gibberellin.

Hormones also play a decisive role in animal growth. One hormone from the pituitary gland at the base of the brain is called growth hormone because of its extensive and widespread effects on growth. A deficiency of growth hormone in pre-adolescents results in dwarfism, and oversupply of the hormone (often caused by a tumour) results in gigantism. If an excess of growth hormone is produced after the long bones can no longer grow—*i.e.,* post-adolescence—a disease called acromegaly, which is characterized by increases in the size of the hands and feet and broadening of facial features, results. A deficiency of thyroid hormone in children also causes growth retardation.

The sex hormones secreted from the pituitary gland interact in a complex way to regulate the growth of the gonads. The gonads in turn produce estrogen and progesterone in females and testosterone in males; these hormones control the development of human secondary sexual characteristics—body hair, enlargement of mammary glands in females, and growth of the vocal cords in males. Although the growth hormones and sex hormones play a vital role in growth, the exact mechanism by which they function has not been established with certainty (see BIOCHEMICAL COMPONENTS OF ORGANISMS: *Hormones*).

In addition to having the ability to synthesize the factors that regulate growth, plants and animals evidently possess exquisite mechanisms for integrating and regulating the production of hormones; *i.e.,* the appropriate amounts of the right hormones are produced at the right time and the right place for normal growth.

Limits of growth

Although many plants, including trees, grow throughout their lives, growth of parts of the organism is not perpetual; *e.g.,* leaves of a given species attain a specific size and can grow no larger. In animals, growth stops entirely, except for replacement, after the juvenile period. The limits for both total body size and organ size are probably established by genetic mechanisms. The factors involved in limiting the growth of an organism are not yet definitely known, but evidence indicates that the liver releases into the bloodstream protein molecules that can limit growth of the organ. Thus, one theoretical view is that an organ may produce substances that serve to limit its own growth, thereby establishing a feedback mechanism. A protein called nerve-growth factor is important for the growth of some parts of the mammalian nervous system. If too much of the nerve-growth factor is present, growth of sympathetic nerve fibres is extensive and aberrant. If the nerve-growth factor is eliminated from the body—by injection of an antibody against the factor—the sympathetic nerves wither and disappear. Other subtle growth regulatory substances specific for various organ systems may eventually be discovered.

The dynamics of growth

MEASUREMENT OF GROWTH

The mathematical analysis of the rate of growth has been a subject of interest for many years. It is based on the rule of cell division: one cell gives rise to two daughter cells. Hence, the theoretical increase in cell number would be a geometric series, in which one cell produces two cells, then four, eight, 16, and so on. In reality, however, the rate of growth is not constant but declines after a period of time, usually because of influences in the environment or because of inherent genetic limitations. Thus the curve showing the growth of cell populations and of organisms is usually S-shaped, or sigmoid, when growth is plotted against time on a graph (see Figure 1). The increase in

From B.I. Balinsky, *An Introduction to Embryology* (1970), W.B. Saunders Company

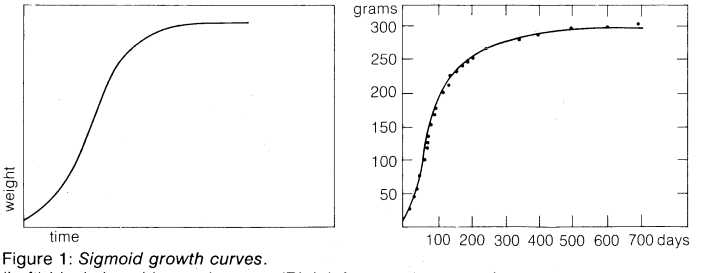


Figure 1: *Sigmoid growth curves.*
(Left) Ideal sigmoid growth curve. (Right) An actual postnatal growth curve of the white rat.

cell number resulting from cell division accounts for the rising part of the curve; the rate of cell division decreases at the plateau in the curve. The S-shaped growth curve is generally applicable to the growth of organisms. If growth is plotted against time on a logarithmic scale, the early intense growth (called log growth) in the rising phase of the growth curve falls on a straight line.

Rate of growth equation

The rate of growth may be defined by the differential equation $v = dW/dt\ (1/W)$, in which v is the growth rate and W is the weight at any given time, t. The solution of this equation provides a value for relative increase—the increase in weight related to the initial mass of the growing substance. The animal that most closely approaches a constant rate of growth is an insect larva. In most animals the rate of growth declines as the organism becomes larger and older.

Although the S-shaped growth curve describes with fair accuracy the growth of populations of single cells, such as bacteria or cells of higher organisms in tissue culture—the growth in a sterile nutrient environment of cells of tissues from organisms—the growth rates of different parts of whole organisms vary. The relationship of the growth of one part of an organism to that in another part is called allometry. An equation expressing the fundamental relationship of allometric growth is $y = bx^k$ in which y is the size of one organ; x is the size of another; b is a constant; and k is known as the growth ratio. Although such mathematical tools have allowed a very thorough

description of the differential growth of different parts of an organism, they have unfortunately not provided insight into the physical and chemical control of the growth rate.

THE STUDY OF GROWTH

Even though the chemical, physical, and genetic bases of growth are elusive, much has been learned about the process by growing tissues in a sterile nutrient environment. Even if the source of the tissue is an organ that has completely stopped growing, such as the nervous system of an animal or the phloem of a plant, the cells will begin to grow again in culture, often at a logarithmic rate of increase. It may therefore be concluded that the organism as a whole places constraints upon the ability of individual cells to reproduce and that, when these constraints are removed, the growth potential of the cells is no longer restrained. Even in tissue culture, however, the rate of cell growth eventually slows, hence the sigmoid-shaped growth curve. During the rapid growth phase of cells in tissue culture, they usually lose the ability to carry out the specialized function characteristic of their organ of origin; for example, if cartilage cells divide rapidly, they no longer synthesize cartilaginous matrix. This phenomenon of apparent despecialization has been a topic of great theoretical interest: are rapid growth and specialization mutually exclusive activities? Evidence shows that some types of specialized cells may be maintained in tissue culture for very long periods of time and still retain the ability to carry out specialized biosyntheses, so that the apparent loss of specialized function in tissue culture cells may not fundamentally result from a mutual exclusivity of growth and differentiation.

Contact inhibition of growth

When the growth of tissue-culture cells begins to slow, one factor responsible is exhaustion of critical components from the medium. But even if the medium is frequently replaced, when the bottom of the culture dish becomes densely packed with a layer of cells, the growth rate drops—a phenomenon called contact inhibition of growth. It is believed that cells so close that they are always touching provide a signal that retards the rate of cell division. Apparently identical cells in tissue culture also show great variation in growth rate. Some cells from the skin, for instance, when placed in culture, may divide every eight hours; other similar cells may divide only every 36 hours. The growth of cells in a controlled environment such as tissue culture offers many possibilities for studying the fundamental mechanisms controlling cell growth and, consequently, the growth of organisms and populations. (F.H.W.)

Aberrations of growth: biological malformation

The processes of development are regulated in such a way that few malformed organisms are found. Those that do appear may, when properly studied, shed light on normal development. The science of teratology—a branch of morphology or embryology—is concerned with the study of these structural deviations from the normal, whether in animals or plants.

In general, abnormalities can be traced to deviations from the normal course of development, often in very early embryonic stages. Such deviations may be caused by abnormal (mutant) genes, by environmental conditions, by infection, by drugs, and, perhaps most frequently, by interactions between these sets of causes. A general interpretation has been that one factor in many cases is reduction of the rate of development, the kind and degree of deformity depending upon the stage at which the retardation occurs. This interpretation is supported by the results of descriptive studies of anomalies, and especially by evidence from experimental teratology. (L.C.D.)

PLANT MALFORMATIONS

Monstrosities, freaks, and other malformations have interested botanists for many years. There are numerous categories of such growth abnormalities in plants, and these are often related only loosely or not at all to one another.

Exaggerated growth. Sometimes divergence from the normal represents merely a quantitative change, which is evidenced by a harmonious but exaggerated manifestation of the normal developmental processes. This is well illustrated in the so-called *bakanae,* or foolish seedling disease, of rice. The *bakanae* disease is caused by the fungus *Gibberella fujikuroi.* Diseased plants are often conspicuous in a field because of their extreme height and pale, spindly appearance. This exaggerated growth response was found to be due to specific substances, known as gibberellins, which were produced by the fungus. Evidence is now available to indicate that gibberellins, also produced by higher plant species, participate directly as an essential growth-regulating system in all higher plant species. The gibberellins of either fungal or higher plant origin stimulate the normal development of certain genetic dwarfs of maize and peas, which cannot themselves produce the gibberellins in amounts sufficient for their normal development.

Effect of the gibberellins

A common deformity of tobacco, called frenching, occurs in most tobacco-growing regions of the world. The advanced state of this condition is characterized by a cessation of terminal bud and stem growth. When dominance of the stem tips is lost, the buds in the axils of the leaves develop, and unusually large numbers of leaves (as many as 300) appear on a plant. The leaves are characteristically sword- or string-shaped because of the failure of the leaf blades to develop. Such plants have the apperance of a rosette. Although the cause of frenching has not yet been unequivocally established, it is thought to be due to a toxic substance produced by the nonpathogenic soil bacterium *Bacillus cereus.*

Alteration of floral parts. Under the stimulus of pathogenic organisms of the most diverse kinds, the sepals, petals, stamens, or pistils of a flower may be transformed into structures that are very different in appearance from those found normally. Certain viruses can cause enlargement of the leaflike flower parts (sepals) surrounding the base of a blossom in plants of the nightshade family. The tomato big-bud virus appears to affect the sepals of the tomato flower rather specifically. These structures enlarge greatly under the influence of the virus and fuse to form huge bladderlike structures that may be 10 times or more the normal size. In the Madagascar periwinkle (*Vinca rosea*), however, viruses of this type bring about a green colouring in the petals, stamens, and styles; normally the petals are pink and the stamens and styles whitish. There is in this instance a retrograde development of floral parts into foliage leaves. (Findings such as these are of interest to the morphologist because they support the contention that the flower should be regarded as a modified leafy branch.)

Translocation of organs. Plant organs may arise in unusual places as a result of the infection by certain types of pathogenic agents. The carrot-yellows virus, for example, stimulates production of aerial tubers in the axils of the leaves of potato plants. Large numbers of adventitious roots (arising in abnormal places) appear on the stems of tomato plants infected with the bacteria *Pseudomonas solanacearum* and *Agrobacterium tumefaciens* as well as the *Fusarium* wilt fungus and the cranberry false blossom virus.

An extreme example of adventitious shoot formation is found in *Begonia phyllomaniaca* after shock. In this instance, small plantlets develop spontaneously in incredible numbers from the superficial cell layers of the leaf blades, petioles, and stems. The adventitious shoots do not arise from preformed buds but develop from cells at the base of hairs and especially from certain glands present in great numbers in young stems and leaves of this species. Although these plantlets develop a vascular system of their own, the vast majority never succeed in connecting that system with the vascular system of the host. They must therefore be regarded not as branches but rather as independent organisms.

The independence of abnormal growths

Witches'-brooms. Witches'-brooms, or hexenbesens, are closely grouped, many-branched structures commonly found on a number of species of trees and shrubs and caused by certain fungi. Witches'-brooms live a more or less independent existence, despite the fact that they are derived from the tissues of the host. In accordance with their independence, the witches'-brooms tend to break

away from the normal correlations of the parent plant. Instead of branching out horizontally, the brooms stand as more or less erect clusters of branches. Witches'-brooms do not as a rule flower, and the vegetative buds may open several weeks earlier in the spring than do those present on healthy branches, indicating further the independence of these structures from normal controls.

Similar structures occur in certain plant species after virus infection. These appear to result from the excessive stimulation and development of secondary shoots. The witches'-broom virus in potato, for example, causes the infected plant to produce numerous buds on the aboveground stems of potato plants. Long, slender stolons resembling aerial roots that are covered with hairs develop from these adventitious buds.

Fasciation. This condition is best placed in that category of teratological abnormalities known as monstrosities. Fasciation is a term that has been used to describe a series of abnormal growth phenomena resulting from many different causes, all of which result in flattening of the main axis of the plant. Although a ribbonlike expansion of the stem is often the most striking feature of this condition, all parts of the plant may be affected. As fasciation develops, the growing point of the plant becomes broader; the unregulated tissue growth results in significant increases in the weight and volume of the plant. The apical growing point becomes linear and comblike in some instances or develops numerous growing points, producing a witches'-broom effect. In still other instances, the growing points may be coiled and resemble a ram's horn, or they may be fused and highly distorted into a grotesque tangle of coils. Fasciations found in plants such as the common cockscomb (*Celosia argentea cristata*) and in cacti are highly prized by gardeners.

There seems little doubt that nutritional changes due to disturbances in the growth-hormone relationships in a plant play an important role in fasciation. It has been suggested that maldistribution of growth hormones in the plant is also a cause of these abnormalities. (A.C.Br.)

ANIMAL MALFORMATIONS

Among the newborn young and embryos of man and most other species of animals are found occasional individuals who are malformed in whole or in part. The most grossly abnormal of these have been referred to from ancient times as monsters, probably because the birth of one was thought to signify something monstrous or portentous; the less severely malformed are usually known as abnormalities or anomalies.

The origin of malformations

Monsters have been regarded by primitive peoples as of supernatural origin. The birth of a malformed individual was often attributed, before the rise of modern science, to intercourse between human beings and devils or animals. The mythical beings that appear in the folklore of many peoples—races of dwarfs and giants, of sirens, mermaids, and men with a single median eye (cyclops) or leg (skiapods)—were probably suggested by observations of malformed humans. Giants and dwarfs were often classed as monsters, probably because of the prominent places they occupied in mythology.

The objective study of malformations began with the English physiologist William Harvey (1651), who correctly attributed them to deviations from the normal course of embryonic development. Systematic scientific study, however, had to await the pioneer work of the French anatomists Étienne and Isidore Geoffroy Saint-Hilaire. Their *Traité de Teratologie* (1836), which laid the basis for the science of teratology, still remains a valuable source of information. Recent improvements in understanding have come from the application of experimental analytical methods and from increased knowledge of the mechanisms of inheritance—*e.g.*, from genetics.

In man certain gross defects in babies at birth have been shown to be associated with effects acting through the mother in early pregnancy: gross defects of eyes and ears caused by infection of the mother with rubella (German measles); and microcephalic idiocy with diagnostic use of X-ray on the mother. The latter observations have been confirmed by animal experimentation. It has been established that higher rates of congenital malformations occur in areas of higher natural radioactivity. The effect may be induced by fallout from atomic explosions.

The teratogenic action of many drugs has been tragically dramatized. The appearance of an alarming number of deformed, basically limbless infants, especially in Germany, in the late 1950s and early 1960s, was traced to the ingestion by pregnant women of the sedative thalidomide (known under many trade names). This drug adversely influences the developing fetus; it appears to interfere with development only in the first seven weeks of pregnancy. Hallucinogenic drugs, such as LSD (lysergic acid diethylamide) are suspected of damaging chromosomes and their use could result in defective offspring.

Thalidomide effects

According to form, two main classes of malformations may be recognized: those with defective or excessive growth in a single body, and those with partial or complete doubling of the body on one of its axes.

Repetition or deficiency of parts. *Somatic characters.* Repetition or deficiency of single parts, such as fingers or toes (polydactyly, hypodactyly, brachydactyly), is a frequent anomaly in man and other mammals. In many analyzed cases it has been shown to result from the inheritance of an abnormal gene that produces a localized disturbance of a growth process in the embryo. In the rabbit a recessive gene for brachydactyly (short digits) causes a localized breakdown of circulation in the developing limb bud of the embryo, followed by necrosis (tissue death) and healing.

Absence or abnormality of whole limbs is less common and includes, besides clubfoot, the so-called congenital amputations once thought to be caused by the strangulation of a limb by a fold of embryonic membrane (amnion). It is probable that internal abnormalities of the bone are more frequent causes of such amputations than are strangulations. Cases are recorded of human identical twins in which both members have the same type of limb abnormality, suggesting a hereditary predisposition to this type of malformation. Besides malformed individuals with rudimentary limbs (phocomelus; having "seal-like limbs"), others have incomplete or underdeveloped extremities (hemimelus, micromelus, ectromelus).

A rare type of malformation, but one that has always attracted special interest, occurs when the lower extremities are more or less united, as in the mythical figures of sirens or mermaids. Such sirenoid individuals may have a single foot (uromelus), or limbs fused throughout their length with no separate feet (sirenomelus or symmelus).

Defects of the brain and head of man

Absence of the brain at birth (anencephaly); an abnormally small brain and head (microcephaly); and enlargement of the brain and head, sometimes to prodigious dimensions due to dilation of the ventricles by fluid (hydrocephaly), are frequent congenital defects in man. In some cases they have been traced to defective genes, although they may also arise from accidental or traumatic processes during embryonic development. Occasionally, malformed persons are found in which a part of the brain protrudes through the cranium as an encephalocoele. An extreme variant of this type is pseudencephaly, in which the whole brain is everted and rests upon the top of the cranium like a wig.

Cyclopian malformations with a single median eye occur rarely in man and other animals. More frequent anomalies are anophthalmia (absence of eyes) and microphthalmia (abnormally small eyes), both occasionally the result of abnormal heredity. Defective closure of lines of junction in the embryo produces malformations such as cleft palate, in which the ventral laminae of the palate have failed to fuse, and harelip, in which the median nasal and maxillary processes fail to unite. A frequent abnormality in human infants is spina bifida, in which the spine fails to close over and a gap is left in the vertebral column. These conditions are inherited, albeit somewhat irregularly, in man.

Sexual anomalies. In man and other vertebrates, male and female individuals usually have distinctive characters in addition to the primary one of producing either sperm or eggs. Individuals with both male and female functions are known as hermaphrodites. While this is the normal condition in some lower animals and in many flowering

plants, it is so rare in mammals as to be regarded as anomalous. Individuals with mixtures of male and female characters (usually sterile) are known as intersexes. In man there occur two rare conditions that, according to recent evidence, represent partial sex reversal. Individuals with Klinefelter's syndrome are apparent males who produce no sperm. Many cases have been shown to have two X-chromosomes (the usual state determining femaleness) with an additional Y-chromosome (which carries genetic factors for maleness). Individuals with Turner's syndrome are apparent females without functional ovaries. The cases analyzed have only one X-chromosome (like the normal male with one X- and one Y-chromosome). These anomalies are clearly caused by disturbances in the mechanism for sex determination.

Complex syndromes. A remarkable feature of malformations in vertebrates including man is the association of multiple abnormalities in complex syndromes. Thus, in man harelip, spina bifida, hydrocephalus, and polydactyly may be found in the same individual; acrocephalosyndactyly (an egg- or dome-shaped skull and partial or complete fusion of digits in both hands and feet) often occurs with harelip, contractures, spina bifida, and mental abnormalities.

In man, individuals afflicted with mongolism, also known as Down's syndrome, have facial and bodily characters that permit diagnosis at or even before birth. Mongols have 47 instead of the normal 46 chromosomes. The extra chromosome is apparently responsible for the abnormal condition.

Doubling of parts. Individuals partially or wholly double, but joined together, are represented by the rare occurrence in man of Siamese twins, so-called from a famous Siamese pair exhibited for many years in the 19th century. The condition consists of identical twins joined by a bridge of tissue through which the circulatory systems communicate. Such twins probably arise by the incomplete separation of a single fertilized egg into two parts; the experimental production of such double individuals in newts has been accomplished by constricting the egg in the two-cell stage.

Siamese twins

In man, partially double symmetrical malformations are found. They vary from those with a single head but with neck, trunk, and limbs doubled, through those with two heads and a single trunk, to others with head, shoulders, and arms doubled, but with one trunk and one pair of legs. Such double malformations probably arise following the less complete separation of the halves of the early embryo or partial separation at later stages. A rare type is one in which there is a Janus head, two faces on a single head and body. Janus malformations have been produced experimentally in amphibian embryos by a variety of treatments in early stages. A group of cases in which the hinder end of the body was doubled from the sacrum back has been found in one strain of mice and appears to be due to abnormal heredity. Doubling of whole limbs in amphibia has been produced experimentally by injuring the limb rudiment at an early, sensitive stage. (For additional information on abnormal human development, see below *Human growth and development: abnormal development.*) (L.C.D.)

Biological regeneration

Organisms differ markedly in their ability to replace lost or amputated body parts. Some grow a new structure on the stump of the old one. By such regeneration whole organisms may dramatically replace substantial portions of themselves when they have been cut in two, or may grow organs or appendages that have been lost. Not all living things regenerate parts in this manner, however. The stump of an amputated structure may simply heal over without replacement. This wound healing is itself a kind of regeneration at the tissue level of organization: a cut surface heals over, a bone fracture knits, and cells replace themselves as the need arises.

Regeneration, as one aspect of the general process of growth, is a primary attribute of all living systems. Without it there could be no life, for the very maintenance of an organism depends upon the incessant turnover by which all tissues and organs constantly renew themselves. In some cases rather substantial quantities of tissues are replaced from time to time, as in the successive production of follicles in the ovary or the molting and replacement of hairs and feathers. More commonly, the turnover is expressed at the cellular level. In mammalian skin the epidermal cells produced in the basal layer may take several weeks to reach the outer surface and be sloughed off. In the lining of the intestines, the life span of an individual epithelial cell may be only a few days.

Cellular renewal

The motile, hairlike cilia and flagella of single-celled organisms are capable of regenerating themselves within an hour or two after amputation. Even in nerve cells, which cannot divide, there is an endless flow of cytoplasm from the cell body out into the nerve fibres themselves. New molecules are continuously being generated and degraded with turnover times measured in minutes or hours in the case of some enzymes, or several weeks as in the case of muscle proteins. (Evidently, the only molecule exempt from this inexorable turnover is deoxyribonucleic acid [DNA] which ultimately governs all life processes.)

There is a close correlation between regeneration and generation. The methods by which organisms reproduce themselves have much in common with regenerative processes. Vegetative reproduction, which occurs commonly in plants and occasionally in lower animals, is a process by which whole new organisms may be produced from fractions of parent organisms; *e.g.,* when a new plant develops from a cut portion of another plant, or when certain worms reproduce by splitting in two, each half then growing what was left behind. More commonly, of course, reproduction is achieved sexually by the union of an egg and sperm. Here is a case in which an entire organism develops from a single cell, the fertilized egg, or zygote. This remarkable event, which occurs in all organisms that reproduce sexually, testifies to the universality of regenerative processes. During the course of evolution the regenerative potential has not changed, but only the levels of organization at which it is expressed. If regeneration is an adaptive trait, it would be expected to occur more commonly among organisms that appear to have the greatest need of such a capability, either because the hazard of injury is great or the benefit to be gained is great. The actual distribution of regeneration among living things, however, seems at first glance to be a rather fortuitous one. It is difficult indeed to understand why some flatworms are able to regenerate heads and tails from any level of amputation, while other species can regenerate in only one direction or are unable to regenerate at all. Why do leeches fail to regenerate, while their close relatives, the earthworms, are so facile at replacing lost parts? Certain species of insects regularly grow back missing legs, but many others are totally lacking in this capacity. Virtually all modern bony fishes can regenerate amputated fins, but the cartilaginous fishes (including the sharks and rays) are unable to do so. Among the amphibians, salamanders regularly regenerate their legs, which are not very useful for movement in their aquatic environment, while frogs and toads, which are so much more dependent on their legs, are nevertheless unable to replace them. If natural selection operates on the principle of efficiency, then it is difficult to explain these many inconsistencies.

Whether regeneration is adaptive

Some cases are so clearly adaptive that there have evolved not only mechanisms for regeneration, but mechanisms for self-amputation, as if to exploit the regenerative capability. The process of losing a body part spontaneously is called autotomy. The division of a protozoan into two cells and the splitting of a worm into two halves may be regarded as cases of autotomy. Some colonial marine animals called hydroids shed their upper portions periodically. Many insects and crustaceans will spontaneously drop a leg or claw if it is pinched or injured. Lizards are famous for their ability to release their tails. Even the shedding of antlers by deer may be classified as an example of autotomy. In all these cases autotomy occurs at a predetermined point of breakage. It would seem that wherever nature contrives to lose a part voluntarily, it provides the capacity for replacement.

Compensatory hypertrophy

Sometimes, when part of a given tissue or organ is removed, no attempt is made to regenerate the lost structures. Instead, that which remains behind grows larger. Like regeneration, this phenomenon—known as compensatory hypertrophy—can take place only if some portion of the original structure is left to react to the loss. If three-quarters of the human liver is removed, for example, the remaining fraction enlarges to a mass equivalent to the original organ. The missing lobes of the liver are not themselves replaced, but the residual ones grow as large as necessary in order to restore the original function of the organ. Other mammalian organs exhibit similar reactions. The kidney, pancreas, thyroid, adrenal glands, gonads, and lungs compensate in varying degrees for reductions in mass by enlargement of the remaining parts.

It is not invariably necessary for the regenerating tissue to be derived from a remnant of the original tissue. Through a process called metaplasia, one tissue can be converted to another. In the case of lens regeneration in certain amphibians, in response to the loss of the original lens from the eye, a new lens develops from the tissues at the edge of the iris on the upper margin of the pupil. These cells of the iris, which normally contain pigment granules, lose their colour, proliferate rapidly, and collect into a spherical mass which differentiates into a new lens.

MODES OF REGENERATION

Basic patterns. Not all organisms regenerate in the same way. In plants and in coelenterates such as the hydra and jellyfishes, missing parts are replaced by reorganization of preexisting ones. The wound is healed, and the neighbouring tissues reorganize themselves into whatever parts may have been cut off. This process of reorganization, called morphallaxis, is the most efficient way for simple organisms to regenerate. Higher animals, with more complex bodies, regenerate parts differently, usually by the production of a specialized bud, or blastema, at the site of amputation. The blastema, made up of cells that look very much alike despite their often diverse origins, made its first appearance evolutionarily in flatworms and is encountered in the regenerative processes of all higher animals. It provides the tissue that will form the regenerated part.

Atypical regeneration. Sometimes the part that grows back is not the same as that which was lost, and, occasionally, regeneration may be induced without having lost anything at all. It is not uncommon for a regenerated part to be incomplete. Earthworms, for example, usually regenerate only five segments in the anterior direction even if more than that number have been amputated. Many insects regenerate abnormally small legs from which some segments may be missing. Tadpole tails when amputated grow back to about only half their original length. These and other cases testify to the fact that a little regeneration is often good enough—that it is not necessary in every case to reproduce a flawless copy of the original.

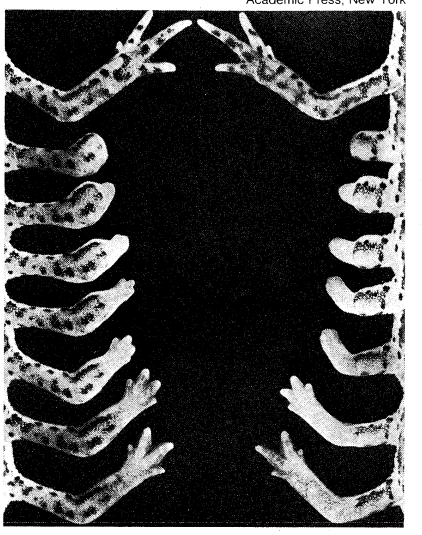
Sometimes that which is regenerated is very different from the original. Among the arthropods there are cases in which the stump of an antenna grows a leg, while a cut eyestalk regenerates an antenna. More commonly, the regenerated part may be a reasonable facsimile of the original but will differ in details. A regenerated lizard tail contains an unsegmented cartilaginous tube instead of a series of vertebrae as did the original tail. The spinal cord lacks segmented ganglia, and the scales in the skin differ in character from the original ones. A regenerated tail, therefore, is easily distinguished from an original one yet appears sufficient to serve the purpose. Another interesting case is that of jaw regeneration in salamanders. If the lower jaw is amputated a new one will grow back, but it is often smaller than the original. It contains teeth and a mandible, but lacks a new tongue. Furthermore, the new mandible is a cartilaginous model of the original, and is not known to convert into bone.

Sometimes more of a part grows back than has been removed by amputation. A limb stump, for example, can occasionally give rise to hands with extra digits. Lobsters have been known to regenerate double structures, in which case the new parts are mirror images of each other.

THE REGENERATION PROCESS

Origin of regeneration material. Following amputation, an appendage capable of regeneration develops a blastema from tissues in the stump just behind the level of amputation (see photograph). These tissues undergo drastic changes. Their cells, once specialized as muscle, bone, or cartilage, lose the characteristics by which they are normally identified (dedifferentiation); they then begin to migrate toward, and accumulate beneath, the wound epidermis, forming a rounded bud (blastema) that bulges out from the stump. Cells nearest the tip of the bud continue

R.J. Goss, *Principles of Regeneration* (1969), Academic Press, New York



Successive stages in the regeneration of opposite limbs in a newt following amputation through (left) lower and (right) upper arms. At the top are the original limbs. From top to bottom are the successive stages of regeneration at 7, 21, 25, 28, 32, 42, and 70 days after amputation. The (right) more proximal stump elongates faster but differentiates more slowly than does the (left) more distal stump.

to multiply, while those situated closest to the old tissues of the stump differentiate into muscle or cartilage, depending upon their location. Development continues until the final structures at the tip of the regenerated appendage are differentiated, and all the proliferating cells are used up in the process.

The blastema cells seem to differentiate into the same kind of cells they were before, or into closely related types. Cells may perhaps change their roles under certain conditions, but apparently rarely do so. If a limb blastema is transplanted to the back of the same animal, it may continue its development into a limb. Similarly, a tail blastema transplanted elsewhere on the body will become a tail. Thus, the cells of a blastema seem to bear the indelible stamp of the appendage from which they were produced and into which they are destined to develop. If a tail blastema is transplanted to the stump of a limb, however, the structure that regenerates will be a composite of the two appendages.

Polarity and gradient theory. Each living thing exhibits polarity, one example of which is the differentiation of an organism into a head, or forward part, and a tail, or hind part. Regenerating parts are no exception; they exhibit polarity by always growing in a distal direction (away from the main part of the body). Among the lower invertebrates, however, the distinction between proximal (near, or toward the body) and distal is not always clear cut. It is not difficult, for example, to reverse the polarity of "stems" in colonial hydroids. Normally a piece of the stem will grow a head end, or hydranth, at its free, or distal, end; if that is tied off, however, it regenerates a hydranth at the end that was originally proximal. The polarity in this system is apparently determined by an activity gradient in such a way that a hydranth regenerates wherever the metabolic rate is highest. Once a hydranth

has begun to develop, it inhibits the production of others proximal to it by the diffusion of an inhibitory substance downward along the stem.

Differences between whole organisms and appendages

When planarian flatworms are cut in half, each piece grows back the end that is missing. Cells in essentially identical regions of the body where the cut was made form blastemas, which, in one case gives rise to a head and in the other becomes a tail. What each blastema regenerates depends entirely on whether it is on a front piece or a hind piece of flatworm: the real difference between the two pieces may be established by metabolic differentials. If a transverse piece of a flatworm is cut very thin—too narrow for an effective metabolic gradient to be set up—it may regenerate two heads, one at either end. If the metabolic activity at the anterior end of a flatworm is artificially reduced by exposure to certain drugs, then the former posterior end of the worm may develop a head.

Appendage regeneration poses a different problem from that of whole organisms. The fin of a fish and the limb of a salamander have proximal and distal ends. By various manipulations, it is possible to make them regenerate in a proximal direction, however. If a square hole is cut in the fin of a fish, regeneration takes place as expected from the inner margin, but may also occur from the distal edge. In the latter case, the regenerating fin is actually a distal structure except that it happens to be growing in a proximal direction.

Amphibian limbs react in a similar manner. It is possible to graft the hand of a newt to the nearby body wall, and once a sufficient blood flow has been established, to sever the arm between the shoulder and elbow. This creates two stumps, a short one consisting of part of the upper arm, and a longer one made up of the rest of the arm protruding in the wrong direction from the side of the animal. Both stumps regenerate the same thing, namely, everything normally lying distal to the level of amputation, regardless of which way the stump was facing. The reversed arm therefore regenerates a mirror image of itself.

Clearly, when a structure regenerates it can only produce parts that normally lie distal to the level of amputation. The participating cells contain information needed to develop everything "downstream," but can never become more proximal structures. Regeneration, like embryonic development, occurs in a definite sequence.

Regulation of regeneration. There are certain prerequisites without which regeneration cannot occur. First and foremost, there must be a wound, although the original appendage need not have been lost in the process. Second, there must be a source of blastema cells derived from remnants of the original structure or an associated one. Finally, regeneration must be stimulated by some external force. The stimuli often involve the nervous system. An adequate nerve supply is required for the regeneration of fish fins, taste barbels, and amphibian limbs. In the case of many tail regenerations, the spinal cord provides the necessary stimulus. Lens regeneration in salamander eyes depends upon the presence of a retina. Arthropod appendages regenerate in the presence of molting hormones. Protozoan regeneration requires the presence of a nucleus. In case after case, regeneration depends on more than a healed wound and a source of blastema cells. It is often triggered by some physiological stimulus originating elsewhere in the body, a stimulus invariably associated with the very function of the structure to be regenerated. The conclusion is inescapable that regeneration is primarily the recovery of deficient functions rather than simply the replacement of lost structures.

The imperative of need is of further importance in suppressing excess regeneration. To be able to regenerate is to run the risk of regenerating too much or too often. If regeneration did not depend upon a physiological stimulus, such as those mediated by nerves or hormones, there would be no reason why simple wounds should not sprout whole new appendages.

Failure of regenerative processes

It is not known why regeneration fails to occur in many cases, as in the legs of frogs or the limbs and tails of mammals. The nerve supply might be inadequate, for when the number of nerves is artificially increased, regeneration is sometimes induced. This cannot be the whole answer, however, because not all appendages depend on nerves for their regeneration; newt jaws, salamander gills, and deer antlers do not require nerves to regenerate.

Possibly the failure to regenerate relates to the ways in which wounds heal. In higher vertebrates there is a tendency to form thick scar tissue in healing wounds, which may act as a barrier between the epidermis and the underlying tissues of the stump. In the absence of direct contact between these two tissues, the stump may not be able to give rise to the blastema cells required for regeneration.

THE RANGE OF REGENERATIVE CAPABILITY

Virtually no group of organisms lacks the ability to regenerate something. This process, however, is developed to a remarkable degree in lower organisms, such as protists and plants, and even in many invertebrate animals such as earthworms and starfishes. Regeneration is much more restricted in higher organisms such as mammals, in which it is probably incompatible with the evolution of other body features of greater survival value to these complex animals.

Protists and plants. *Algae.* One of the most outstanding feats of regeneration occurs in the single-celled green alga *Acetabularia.* This plant-like protist of shallow tropical water consists of a group of short rootlike appendages; a long thin "stem," up to several centimetres in length; and an umbrella-like cap at the top. The entire organism is one cell, with its single nucleus situated at the base in one of the "roots." If the cap is cut off, a new one regenerates from the healed over stump of the amputated stem. The nucleus is necessary for this kind of regeneration, presumably because it provides the information needed to direct the development of the new cap. Once this information has been produced by the nucleus, however, the nucleus can be removed and regeneration continues unabated.

If the nucleus from one species of *Acetabularia* is added to a cell-body of another species, and the cap of the recipient cell is amputated, the new cap that regenerates will be a hybrid because each nucleus exerts its own morphogenetic influences. On the other hand, if the nucleus from one species is substituted for that in another, regeneration reflects the properties of the new nucleus.

Protozoans. Most single-celled, animal-like protists regenerate very well. If part of the cell fluid, or cytoplasm, is removed from *Amoeba,* it is readily replaced. A similar process occurs in other protozoans, such as flagellates and ciliates. In each case, however, regeneration occurs only from that fragment of the cell containing the nucleus. Amputated parts that lack a nucleus cannot survive. In some ciliates, such as *Blepharisma* or *Stentor,* the nucleus may be elongated or shaped like a string of beads. If either of these organisms is cut in two so that each fragment retains part of the elongated nucleus, each half proceeds to grow back what it lacks, giving rise to a complete organism in less than six hours. The way in which such a bisected protozoan regenerates is almost identical with the way it reproduces by ordinary division. Even a very tiny fragment of the whole organism can regenerate itself, provided it contains some nuclear material to determine what is supposed to be regenerated.

Nuclear material

Green plants. The mechanisms by which vascular plants grow have much in common with regeneration. Their roots and shoots elongate by virtue of the cells in their meristems, the conical growth buds at the tip of each branch. These meristems are capable of indefinite growth, especially in perennial plants. If they are amputated they are not replaced, but other meristems along the stem, normally held in abeyance, begin to sprout into new branches that more than compensate for the loss of the original one. Such a process is called restitution.

Plants are also capable of producing callus tissue wherever they may be injured. This callus is proliferated from cambial cells, which lie beneath the surface of branches and are responsible for their increase in width. When a callus forms, some of its cells may organize into growing points, some of which in turn give rise to roots while others produce stems and leaves.

Callus tissue

Invertebrates. *Coelenterates.* The vast majority of research on coelenterates has been focussed on hydras and

some of the colonial hydroids. If a hydra is cut in half, the head end reconstitutes a new foot, while the basal portion regenerates a new hydranth with mouth and tentacles. This seemingly straightforward process is deceptively simple. From tiny fragments of the organism whole animals can be reconstituted. Even if a hydra is minced and the pieces scrambled, the fragments grow together and reorganize themselves into a complete whole. The indestructibility of the hydra may well be attributed to the fact that even the intact animal is constantly regenerating itself. Just below the mouth is a growth zone from which cells migrate into the tentacles and to the foot where they eventually die. Hence, the hydra is in a ceaseless state of turnover, with the loss of cells at the foot and at the tips of the tentacles being balanced by the production of new ones in the growth zone. If such an animal is X-rayed, the proliferation of new cells is inhibited and the hydra gradually shrinks and eventually dies owing to the inexorable demise of cells and the inability to replace them.

In colonial hydroids, such as *Tubularia,* there is a series of branching stems, each of which bears a hydranth on its end. If these hydranths are amputated they grow back within a few days. In fact, the organism normally sheds its hydranths from time to time and regenerates new ones naturally.

Planarian regeneration

Flatworms. Planarian flatworms are well-known for their ability to regenerate heads and tails from cut ends (see Figure 2). In the case of head regeneration, some

After (top) T.H. Morgan in E. Korschelt, *Regeneration und Transplantation,* from (bottom) V. Hamburger, *Manual of Experimental Embryology* (© 1960), University of Chicago Press

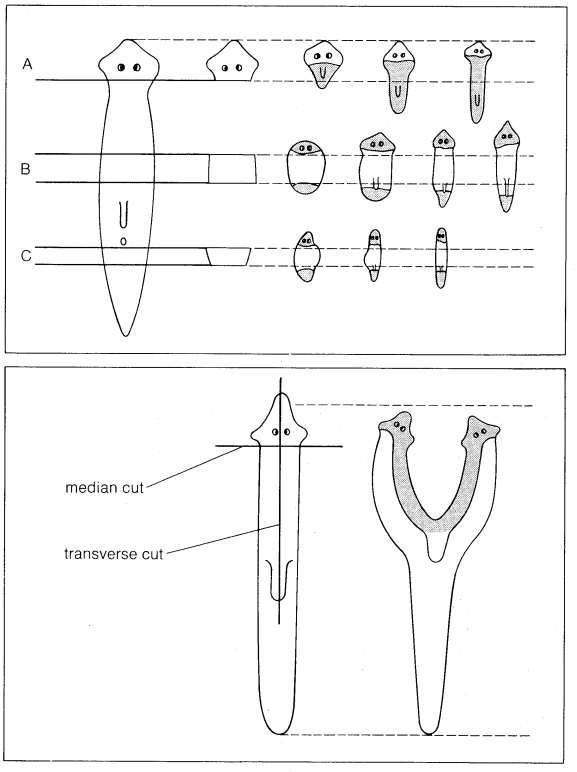


Figure 2: (Top) Regeneration in the planarian flatworm *Dugesia*. The three rows show regeneration from (A) anterior section, (B) midsection, and (C) posterior section of the animal at the left. (Bottom) Regeneration of a double head in a planarian. Regenerated tissue is shaded.

blastema cells become brain tissues, others develop into the eyes, and still others differentiate as muscle or intestine. In a week or so, the new head functions almost as well as the original.

The blastema that normally gives rise to a single head is, under certain circumstances, even capable of becoming two heads if the stump of a decapitated flatworm is divided in two by a longitudinal cut. Each of the two halves then gives rise to a complete head. Thus, each blastema develops into an entire structure regardless of its size or position in relation to the rest of the animal.

In the case of flatworms there is still considerable disagreement concerning the origins of the blastema. Some investigators contend that it is derived from neoblasts, undifferentiated reserve cells scattered throughout the body. Others claim that there are no such reserve cells and that the blastema develops from formerly specialized cells near the wound that dedifferentiate to give rise to the blastema cells. Whatever their source, the cells of the blastema are capable of becoming many different things depending upon their location.

Regeneration in flatworms occurs in a stepwise fashion. The first tissue to differentiate is the brain, which induces the development of eyes. Once the head has formed, it in turn stimulates the production of the pharynx. The latter then induces the development of reproductive organs farther back. Thus, each part is necessary for the successful development of those to come after it; conversely, each part inhibits the production of more of itself. If decapitated flatworms are exposed to extracts of heads, the regeneration of their own heads is prevented. Such a complex interplay of stimulators and inhibitors is responsible for the successful regeneration of an integrated morphological structure.

Annelids. The segmented worms exhibit variable degrees of regeneration. The leeches, as already noted, are wholly lacking in the ability to replace lost segments, whereas the earthworms and various marine annelids (polychaetes) can often regenerate forward and backward. The expression of such regenerative capacities depends very much on the level of amputation. Anteriorly directed regeneration usually occurs best from cuts made through the front end of the worm, with little or no growth taking place from progressively more posterior bisections. Posteriorly directed regeneration is generally more common and extensive. Some species of worms replace the same number of segments as were lost. Hypomeric regeneration, in which fewer segments are produced than were removed, is more common, however.

Anterior regeneration depends upon the presence of the central nerve cord. If this is cut or deflected from the wound surface, little or no forward regeneration may take place. Posterior regeneration requires the presence of the intestine, removal of which precludes the formation of hind segments. Thus, it would seem that no head will regenerate without a central nervous system, nor a tail without an opening.

Arthropods. Many insects and crustaceans regenerate legs, claws, or antennas with apparent ease. When insect legs regenerate, the new growth is not visible externally because it develops within the next proximal segment in the stump. Not until the following molt is it released from its confinement to unfold as a fully developed leg only slightly smaller than the original. In the case of crabs, regenerating legs bulge outward from the amputation stump. They are curled up within a cuticular sheath, not to be extended until the sheath is molted. Lobsters and crayfish regenerate claws and legs in a straightforward manner as direct outgrowths from the stumps. As in other crustaceans, however, these regenerates lie immobile within an enveloping cuticle and do not become functional until their sheath is shed at the next molt.

Molting

In all arthropods regeneration is associated with molting, and therefore takes place only during larval or young stages. Most insects do not initiate leg regeneration unless there remains ample time prior to the next scheduled molt for the new leg to complete its development. If amputation is performed too late in the intermolt period, the onset of regeneration is delayed until after shedding; the regenerate then does not appear until the second molt. Metamorphosis into the adult stage marks the end of molting in insects, and adults accordingly do not regenerate amputated appendages.

Crustaceans often tend to molt and grow throughout life. They therefore never lose the ability to grow back missing appendages. When a leg is lost, a new outgrowth appears even if the animal is not destined to molt for many months. Following a period of basal growth, during which a diminutive limb is produced, the regenerated part eventually ceases to elongate. Not until a few weeks before the

next molt does it resume growth and complete its development, triggered by the hormones that induce molting.

Vertebrates. *Fishes.* Many different parts of the fish's body will grow back. Plucked scales are promptly replaced by new ones, and amputated gill filaments can regenerate easily. The "whiskers," or taste barbels, of the catfish grow back as perfect replicas of the originals. The most conspicuous regenerating structures in fishes, however, are the fins. When any of these are amputated, new fins grow out from the stumps and soon restore everything that was missing. Even the coloured stripes or spots that adorn some fins are reconstituted by new pigment cells that repopulate the regenerated part. Fin regeneration depends on an adequate nerve supply. If the nerves are cut leading into the fin, regeneration of neither the amputated fin nor excised pieces of the bony fin rays can take place.

Fin regeneration

Amphibians. Salamanders are remarkable for their ability to regenerate limbs. Larval frogs, or tadpoles, also possess this ability, but usually lose it when they become frogs. It is not known why frog legs do not regenerate, and under appropriate stimuli they can be induced to do so.

Tadpoles and salamanders can replace amputated tails. Tadpole tails have a stiff rod called the notochord for support, whereas salamanders possess a backbone, composed of vertebrae. Both tails contain a spinal cord. When the salamander regenerates its tail, the spinal cord grows back and segmental nerve-cell clusters (ganglia) differentiate. Tadpoles also regenerate their spinal cords, but not the associated ganglia. If the spinal cord is removed or destroyed in the salamander, no tail regeneration occurs; if it is removed from the tadpole tail, however, regeneration can proceed without it.

Reptiles. Lizards also regenerate their tails, especially in those species that have evolved a mechanism for breaking off the original tail when it is grasped by an enemy. When the lizard tail regenerates, however, it does not replace the segmented vertebrae. Instead, there develops a long tapering cartilaginous tube within which the spinal cord is located and outside of which are segmented muscles. The spinal cord of the lizard tail is necessary for regeneration, but the regenerated tail does not reproduce the ganglia that are normally associated with it. Occasionally, a side tail may be produced if the original tail is broken but not lost.

Birds. Regeneration of amputated appendages in birds is not known to occur; however, they do replace their feathers as a matter of course. While most species shed and regenerate feathers one at a time so as not to be grounded, flightless birds, such as penguins, may molt them all at once. Male puffins cast off their colorful beaks after the mating season, but grow new ones the following year. In like manner, the dorsal keel on the upper beaks of male pelicans is shed and replaced annually.

Mammals. Although mammals are incapable of regenerating limbs and tails, there are a few exceptional cases in which lost tissues are in fact regenerated. Not the least of these cases is the annual replacement of antlers in deer. These remarkable structures, which normally grow on the heads of male deer, consist of an inner core of bone enveloped by a layer of skin and nourished by a copious blood supply. During the growing season the antlers elongate by the proliferation of tissues at their growing tips. The rate of growth in some of the larger species may surpass one centimetre (0.39 inch) per day; the maximum rate of growth recorded for the elk is 2.75 centimetres (1.05 inches) per day. When the antlers have reached their full extent, the blood supply is constricted, and the skin, or velvet, peels off, thus revealing the hard, dead, bony antlers produced by the male deer in time for the autumn mating season. The regeneration of elk antlers spans about seven months (see Figure 3). The following spring, the old antlers are shed and new ones grow to replace them.

Antler replacement

R.J. Goss, *Clinical Orthopaedics and Related Research*, no. 69, p. 228 (1970), J.B. Lippincott Company

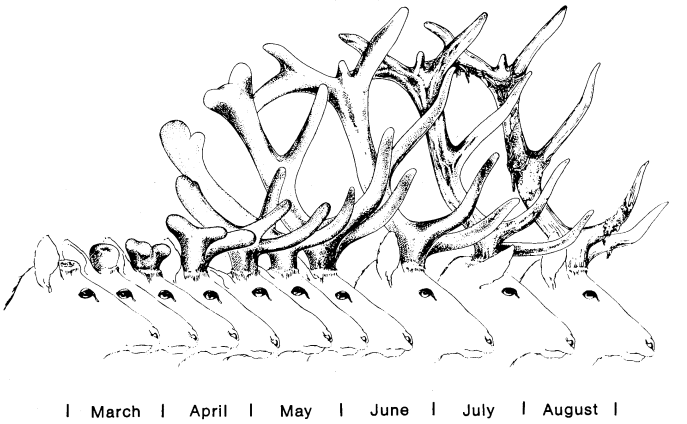


Figure 3: Regeneration of antlers in the elk (see text).

Still another example of mammalian regeneration occurs in the case of the rabbit's ear. When a hole is punched through the external ear of the rabbit, tissue grows in from around the edges until the original opening is reduced or obliterated altogether. This regeneration is achieved by the production of new skin and cartilage from the margins of the original hole. A similar phenomenon occurs in the case of the bat's wing membrane. (R.J.G.)

GENERAL FEATURES OF BIOLOGICAL DEVELOPMENT

Development in its most general meaning refers to any process of progressive change. In this sense, most modern philosophical outlooks would consider that development of some kind or other characterizes all things, in both the physical and biological worlds. Such points of view go back to the very earliest days of philosophy.

Early theories of development

Among the pre-Socratic philosophers of Greek Ionia, half a millennium before Christ, some, like Heracleitus, believed that all natural things are constantly changing. In contrast, others, of whom Democritus is perhaps the prime example, suggested that the world is made up by the changing combinations of atoms, which themselves remain unaltered, not subject to change or development. The early period of post-Renaissance European science may be regarded as dominated by this latter atomistic view, which reached its fullest development in the period between Newton's laws of physics and Dalton's atomic theory of chemistry in the early 19th century. This outlook was never easily reconciled with the observations of biologists, and in the last hundred years a series of discoveries in the physical sciences have combined to swing opinion back toward the Heracleitan emphasis on the importance of process and development. The atom, which seemed so unalterable to Dalton, has proved to be divisible after all, and to maintain its identity only by processes of interaction between a number of component subatomic particles, which themselves must in certain aspects be regarded as processes rather than matter. Albert Einstein's theory of relativity showed that time and space are united in continuum, which implies that all things are involved in time; that is to say, in development.

The philosophers who charted the transition from the nondevelopmental view, for which time was an accidental and inessential element, were Henri Bergson and, in particular, Alfred North Whitehead. Karl Marx and Friedrich Engels, with their insistence on the difference between dialectical and mechanical materialism, may be regarded as other important innovators of this trend, although the generality of their philosophy was somewhat compromised by the political context in which it was placed and the rigidity with which their later followers have interpreted it.

Philosophies of the Heracleitan type, which emphasize process and development, provide much more appropriate frameworks for biology than do philosophies of the atomistic kind. Living organisms confront biologists with changes of various kinds, all of which could be regarded

as in some sense developmental; however, biologists have found it convenient to distinguish the changes and to use the word development for only one of them. Biological development can be defined as the series of progressive, nonrepetitive changes that occur during the life history of an organism. The kernel of this definition is to contrast development with, on the one hand, the essentially repetitive chemical changes involved in the maintenance of the body, which constitute "metabolism," and on the other hand, with the longer term changes, which, while nonrepetitive, involve the sequence of several or many life histories, and which constitute evolution.

'he lurring of istinc- ons in ature

As with most formal definitions, these distinctions cannot always be applied strictly to the real world. In the viruses, for instance, and even in bacteria, it is difficult to make a distinction between metabolism and development, since the metabolic activity of a virus particle consists of little more than the development of new virus particles. In certain other cases, the distinction between development and evolution becomes blurred: the concept of an individual organism with a definite life history may be very difficult to apply in plants that reproduce by vegetative division, the breaking off of a part that can grow into another complete plant. The possibilities for debate that arise in these special cases, however, do not in any way invalidate the general usefulness of the distinctions as conventionally made in biology.

The scope of development

All organisms, including the very simplest, consist of two components, distinguished by a German biologist, August Weismann, at the end of the 19th century, as the "germ plasm" and the "soma." The germ plasm consists of the essential elements, or genes, passed on from one generation to the next, and the soma consists of the body that may be produced as the organism develops (Figure 4). In more modern terms, Weismann's germ plasm is identified with DNA (deoxyribonucleic acid), which carries, encoded in the complex structure of its molecule, the instructions necessary for the synthesis of the other compounds of the organism and their assembly into the appropriate structures. It is this whole collection of other compounds (proteins, fats, carbohydrates, and others) and their arrangement as a metabolically functioning organism that constitutes the soma. Biological development encompasses, therefore, all the processes concerned with implementing the instructions contained in the DNA. Those instructions can only be carried out by an appropriate executive machinery, the first phase of which is provided by the cell that carries the DNA into the next generation: in animals and plants by the fertilized egg cell; in viruses by the cell infected. In life histories that have more than a minimal degree of complexity, the executive machinery itself becomes modified as the genetic instructions are gradually put into operation, and new mechanisms of protein synthesis are brought into functional condition. The fundamental problem of developmental biology is to understand the interplay between the genetic instructions and the mechanisms by which those instructions are carried out.

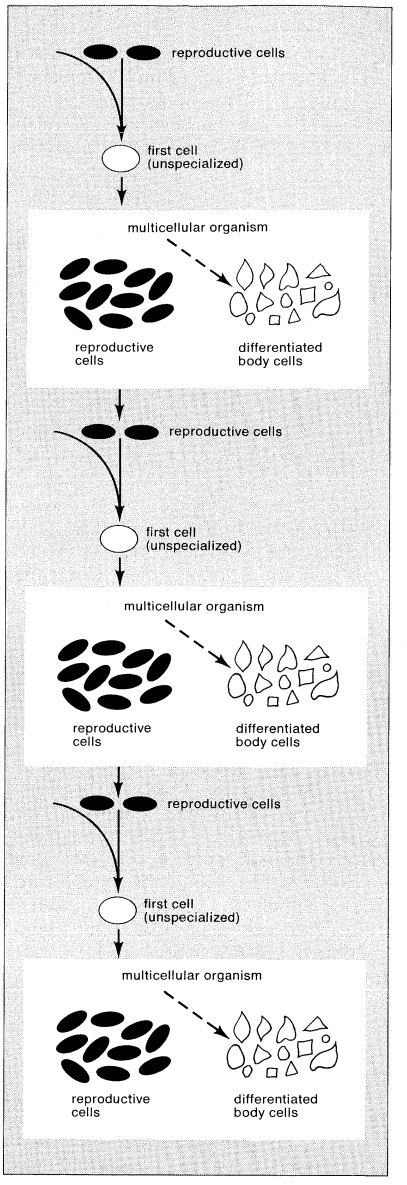


Figure 4: Weismann's concept of the continuity of the germ plasm.

From *Biological Sciences Curriculum Study Green Version High School Biology*, 2nd ed. (1968); Rand McNally & Company

'enotype 'ersus henotype

In the language of genetics the word genotype is used to indicate the hereditary instructions passed on from one generation to another in the genes, while phenotype is the term given to the functioning organisms produced by those instructions. Biological development, therefore, consists of the production of phenotypes. The point made in the last paragraph is that the formation of the phenotype of one generation depends on the functioning of part of the phenotype of the previous generation (*e.g.*, egg cell), as the mechanism that begins the interpretation of the instructions contained in the new organism's genotype.

TYPES OF DEVELOPMENT

In the entire realm of organisms, many different modes of development are found, the most important categories of which can be discussed as pairs of contrasting types.

Quantitative and qualitative development. Development may amount to no more than a quantitative change (usually an increase) in a system that remains essentially unaltered. Qualitative development involves an alteration in the nature of the system. Pure examples of the first type are difficult to find. Approximations to it occur when an animal or plant has attained a structure with the full complement of organs; it then appears to increase only in size, that is to say, quantitatively. This would be a period of simple growth. A closer examination nearly always shows that the system is also undergoing some qualitative change, however. A human infant at birth, for example, already has its full complement of organs, but the ensuing developmental period up to adulthood involves not only growth but also processes of maturation that involve qualitative as well as quantitative changes. Perhaps the most uncomplicated examples of quantitative development occur in certain simple plants and animals. Flatworms, for example, may become reduced in size when starved but increase in size again when provided with suitable nutrition; they thus undergo quantitative changes. Even in these cases, however, it is found that the constituent organs do not always merely become reduced in size but may actually suffer the loss of certain parts.

Progressive and regressive development. The normal processes of development in the majority of plants and animals may be considered progressive since they lead to increases in size and complexity and to the addition of

The adaptability of regression

new elements to the system. As already indicated, some organisms, when placed in adverse conditions, may undergo regressive changes, both in size and complexity. Such regressive changes are a part of the normal life history of certain organisms. Characteristically, these are species in which the organism at an early stage develops a relatively complex structure that enables it to be motile, and later adopts a form of life for which motility is no longer a necessity. A good example is that of the barnacles, a group of marine crustaceans in which the egg at first develops into a motile larva that soon settles down and becomes firmly attached to a solid underwater surface. The barnacle then loses many of the organs characteristic of the motile phase and develops into its familiar stationary form.

There are a number of other examples, particularly in groups in which the adults adopt a parasitic form of life, especially within the digestive system or other tissues of a host animal, from which they have only to absorb their nutriment without having to move or to possess suitable organs for capturing prey. In such cases the early developmental period is characterized by progression toward more complex forms followed by a period of regression in which many of these organs may be lost. During this regressive period certain components of the organism (*i.e.,* those concerned with functioning as a sessile or parasitic form) may undergo progressive development at the same time as the other organs are regressing.

Single-phase and multiphase development. The most familiar organisms, including man, undergo a single-phase development; the organs that appear at early stages persist throughout the whole of life. There are many kinds of animals that develop one or more larval stages adapted to a life different from that of the adult. Perhaps the best known of these is the common frog. The egg first develops into a tadpole, which is provided with a large muscular tail by which it swims. The tadpole eventually undergoes a change of form, or metamorphosis. This involves the regression and resorption of the tail and the growth of the limbs. During this time the rest of the body of the tadpole undergoes less profound changes; the organs persist but undergo relatively far-reaching progressive changes. In other animals, the alteration between the larval and the adult forms may be much more drastic. The egg of a sea urchin, for instance, at first develops to a small larva (the pluteus), which is completely unlike that of the adult. During metamorphosis nearly all the structures of the pluteus disappear; the five-rayed adult develops from a very small rudiment within the larva. In other groups of marine invertebrates, there may be successive larval stages before the adult form appears.

Alternation of generations

Plants in general appear to exhibit a type of development related in a general way to the multiphased development just discussed in animals, although rather different from it in essence. This is called the "alternation of generations" (Figure 5). The majority of higher plants possess two sets of similar chromosomes in each of their cells, that is to say they are diploid ($2n$), as are most higher animals. But in sexual reproduction, diploid cells undergo a reduction division so as to form precursors of the sex cells, which are haploid—*i.e.,* they contain only one set of chromosomes. In animals these cells develop directly into the sex cells—egg and sperm—which unite in fertilization. In plants the haploid cells undergo some developmental processes before the functioning sex cells are produced. The products of this development are spoken of as the "haploid generation." In most higher plants the haploid development is quite reduced, so that the haploid individuals contain only a few nuclei—those associated with the pollen tube on the male side and a few associated with the egg on the female side. In some lower plants, however, such as mosses and ferns, the haploid development may be much more extensive and give rise to quite sizable separate plants. In such cases a species contains two kinds of individuals, produced by different types of developmental processes controlled, however, by the same genotype. This may be compared with the multiphasing development of larval forms in animals. The situation in plants, however, is characterized by the two forms of the organism having different chromosomal constitutions—haploid and diploid—whereas the larval forms and the adult of an animal species have the same chromosomal constitution.

From H.A. Moore and J.R. Carlock, *The Spectrum of Life*; copyright © 1970 by Harper & Row, Publishers, Inc.

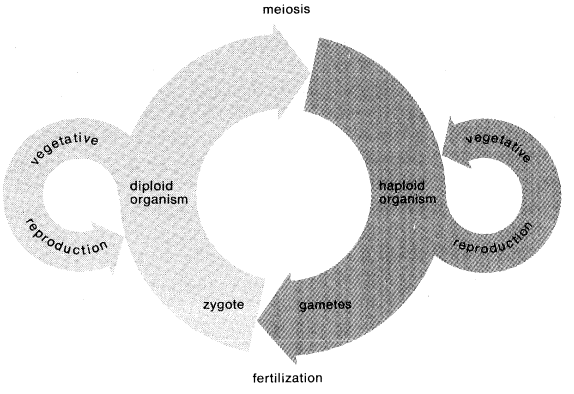


Figure 5: Generalized life cycle showing alternation of generations.

Form and function

Structural and functional development. These two categories cannot be regarded as a pair of opposites as were the previous pairs in this list; rather, they are two aspects of all processes of biological development and can be separated only conceptually, and for purposes of convenience of description. Function is the capacity of the biological system to carry out operations. At the level of the organism, these operations include walking, swimming, eating, digesting, etc.; at the cell level, typical functions are respiring, contracting, conducting nervous impulses, secreting hormones, etc.; and at the molecular level, all functions depend on the production of enzymes, coded by particular genes. Structure encompasses all parts of the organism capable of carrying out functions localized within the body of the organism and arranged in some particular spatial pattern. Contractile cells, for example, are grouped together to form muscle, and other cells are grouped together to form elements of the skeleton; both the muscles and the skeletal elements have definite spatial relations to each other.

These two aspects of development—function and structure—are not opposed to each other in any way. On the contrary, it is obvious that the higher level functions are clearly dependent on the proper structural relations and functions of cell systems. Even at the basic cellular or molecular levels, secretion or nervous conduction essentially depends upon the proper structural relation of the subcellular elements. It is, however, often convenient to focus discussion on one or other of these two aspects of development; for instance, a study may be made into the developmental processes that bring about the production of hemoglobin or insulin by a certain kind of cell, without at the moment being concerned with structural problems. Or again, the focus may be on the results of a certain process by which a mass of cells develops into a typical hand with five digits. In such an inquiry the structural aspects are paramount.

The latitude of normal development

Normal and abnormal development. If a number of fertilized eggs of a given species are provided with conditions that enable them to develop at all, they will, with extraordinary regularity, develop into exceedingly similar adult organisms. The range of conditions they can tolerate is rather wide, and the similarity of the end products surprisingly complete. There are, indeed, good grounds for recognizing what must be considered normal development. The situation is perhaps more marked in animals than in plants, since the plants produced from a given batch of seed under a variety of environmental conditions often present considerably greater variation than is commonly found among animals. Even among plants, however, the differences produced by different conditions of cultivation are usually no more than quantitative differences in size and number of such organs as leaves and flowers, so that an individual can be described as well or poorly developed rather than as normally or abnormally developed. It is only in relatively few cases that a plant develops in quite different ways under two different conditions, neither of

which can be considered abnormal or normal. In certain aquatic plants, for instance, the shape of the emergent leaves is different from the leaves that develop underwater. In such cases the plant actually has two normal forms of development.

It is possible, of course, to produce abnormal organisms by submitting a developing system to stimuli not usually encountered in a normal environment, such as certain chemicals. The presence of unusual genes also may result in deviations from the normal processes of development. In the vast majority of cases such abnormalities can be regarded as resulting from failure to carry out fully the normal processes of development. Functional abnormality in the adult consists in the failure of the system to produce a certain enzyme or functional cell type; a structural abnormality consists in the unusual appearance of certain component elements or in their arrangement in incompletely realized patterns. It is extremely rare to find examples in which the abnormality consists in the addition of a new enzyme not produced in normal development, or the formation of a new structural pattern of the elements.

Cancerous growth as progressive

One very important type of development that, from some points of view, can be considered as an exception to the rule that abnormal development is nearly always retrogressive, is carcinogenesis, the production of tumours. Carcinogenesis involves a change in the developmental behaviour of a group of cells. Initially, it often involves a loss of some of the functional and structural characteristics that previously appeared in the cells. It is commonly followed, however, by the assumption of new properties, which however untoward they may be for the host animal, must be considered as a progressive type: the cells often grow faster and multiply sooner than the noncancerous cells, for example. Furthermore, the cells may undergo a sequence of changes in character and in their arrangement within the tumour. All these features can be regarded in a developmental sense as progressive.

In view of the great rarity of cases of abnormal development that lead to progressive changes, it seems to follow that the organs produced during the normal development of any given species actually exhaust all the potentialities of its genotype for the production of orderly functional structures. It appears that the only abnormal developments that can be produced are either displacements of normal organs, or inadequacies in carrying out normal processes, or the initiation of progressive but quite disorderly processes, as in the production of tumours (see above *Aberrations of growth: biological malformation*).

GENERAL SYSTEMS OF DEVELOPMENT

Development of single-celled organisms. In viruses, activities consist in the production, aided by the machinery of a host cell, of units for building new virus or phage particles: development is simply the assemblage of these constituent units.

In the next higher grade of biological organization, the organism consists of a single cell. Many single-celled algae produce special forms of cells that correspond to the sex cells, or gametes; these cells may unite in fertilization, the resulting fertilized egg, or zygote, undergoing a short period of development. In many other single-celled organisms, however, reproduction takes place by the simple division of an original cell into two daughter cells. In such forms, development normally is part of the process of subdivision. It involves the remodelling of the parent cell into two smaller cells, which are then separated by the division. Something similar must, of course, be involved in the division of cells of higher organisms also. In many single-celled organisms, however, the cell contains a number of defined parts, which are arranged in very definite ways, so that the process of remodelling is very striking and easily observed. This is so, for instance, with ciliated protozoans, in which the cortex is provided with a large number of hairlike cilia or other appendages, arranged in precise patterns, and often with such other structures as a mouth or a gullet. These structures are reproduced in two identical but smaller copies during cell division. This does not necessarily imply that no other developmental processes are possible. The process of regeneration of parts removed occurs quite independently of cell division, for example.

Protozoan development

Open and closed systems of development. There is a marked difference between the general system of development in multicellular plants and multicellular animals. In a plant, certain groups of cells retain throughout the whole life of the plant an embryonic capability to give rise to many types of cells. These regions, known as meristems, occur at the growing tips of branches and roots and as a cylindrical sheath around the stem. They consist of rapidly dividing cells capable of assembling into groups that form buds from which may arise new stems, leaves, flowers, or roots.

By contrast, most animals have no special regions that retain an embryonic character. In most forms, the whole egg, and the whole collection of cells immediately derived from it, take part in the developmental processes and form parts of the developing embryo. In some forms that go through a number of larval stages, the development of certain cells is interrupted at an early stage, and they are set aside and resume their development to form a later type of larva, or to form the adult after the larval stages are completed. An example would be the imaginal buds of some insects. The cells of these buds cannot be regarded as retaining a fully embryonic character comparable to that of the plant meristems, since they cannot perform all the developmental processes but only those involved in the production of the particular late-larval or adult structure for which they have been set aside. In general, then, plants remain embryonic in character, capable as it were of starting again from the beginning to carry out the entire developmental process. Their development is, in this sense, "open." Most animals, on the other hand, lack persistently embryonic cells of this kind, and their development may be characterized as "closed." (There may be certain exceptions to this in very simple forms, such as flatworms, in which certain cells called neoblasts seem able to participate in any type of development; these cells are usually scattered throughout the body, and the major developmental processes that bring into being the general form of the organism cannot be attributed to them, as the development of the plant can be attributed to the meristems.)

Blastogenesis versus embryogenesis. Some animals possess a second system of development, in contrast to the "closed" embryonic system emphasized in the last section. In its most fully developed form, this system consists in remodelling a portion of the parental body into a new organism without any involvement of eggs or sperm. In an adult hydra, a microscopic aquatic animal, a portion of the body may begin to grow exceptionally fast; its cells differentiate into the various cell types and become molded into the constituent organs to build up a new individual identical to the parent. The group of cells responsible for this behaviour is, in its early stages, referred to as a bud, or blastema. Before they become activated these cells may appear quite indistinguishable from the other cells of the body and betray no embryonic capability comparable to the meristems of plants.

In some higher organisms, including certain insects, reptiles, and amphibians, incomplete but still fairly extensive new developments of a similar kind may take place. They require the stimulus of an injury, however, which may involve the removal of part of the normal body. The usual result is a new development to regenerate, or replace, the missing part. The first stage in such regenerative processes consists in the formation of a blastema, that is, a group of rapidly dividing cells that shows little sign of cellular specialization. The evidence indicates that they may not arise, as was once thought, from persisting embryonic cells scattered within the adult body, but instead are formed of cells near the position of the injury. These cells lose their normal adult character and become capable of developing into most of the tissues required to replace the parts removed by the injury (see above *Biological regeneration*).

The developmental aspect of regeneration

Development from a blastema, or blastogenesis, presents many contrasts to embryogenesis, the normal form of development from a fertilized egg. In blastogenesis, tissues that, during embryonic development, appear in se-

quence one after another, may be formed simultaneously and without any obvious sequential relations. Very little, however, is as yet understood about the mechanisms by which the various tissues within the blastema become differentiated from one another. It may well be that these mechanisms are more similar to those found in embryonic development than appears at first sight.

CONSTITUENT PROCESSES OF DEVELOPMENT

Growth. As was pointed out earlier, developing systems normally increase in size, at least during part of their development. "Growth" is a general term used to cover this phenomenon. It comprises two main aspects: (1) increase in cell numbers by cell division and (2) increase in cell size. These two processes may in some examples occur quite separately from each other; for instance, cells in certain rapidly growing tissues (*e.g.,* the connective tissue or blood-forming systems in vertebrates) may increase greatly in number, while the cells remain approximately the same size. Alternatively, in some organs (*e.g.,* the salivary glands of insects) the cells may increase greatly while remaining the same in number, each cell becoming enlarged, or hypertrophied. In such greatly enlarged cells there is often duplication of the genes, involving an increase in the DNA content of the nucleus, although no cell division takes place, and the nucleus continues as a single body, although with a multiplied, or "polyploid," set of chromosomes.

Size and number

In very many cases, however, the growth of an organ depends on increases both in cell number and in cell size. The relative importance of these two processes has yet to be properly investigated. One case that has been well studied is the size of the wings of the fruit fly *Drosophila.* The number of cells in the wing can be easily determined, since each bears a single hair that can be seen and counted in simple microscopic preparations. It has been found that there is an accommodation of factors: if there is an unusually large number of cells, these may be somewhat smaller than usual, so that the total size of the wing remains relatively unchanged.

Perhaps the major theoretical difficulty in the concept of growth is that it is a quantitative notion attached to an ill-defined entity. Growth is an increase in size; but size of what? If a cell or organ increases in volume merely by the absorption of water, or by the laying down of a mineral substance such as calcium carbonate, is this to be regarded as growth or not?

Morphogenesis. As was pointed out earlier, morphogenesis refers to all those processes by which parts of a developing system come to have a definite shape or to occupy particular relative positions in space. It may be regarded as the architecture of development. Morphogenetic processes involve the movement of parts of the developing system from one place to another in space, and therefore involve the action of physical forces, in contrast to processes of differentiation (see below), which require only chemical operations. Although in practice the physical and chemical processes of development normally proceed in close connection, for purposes of discussion it is often convenient to make an artificial separation between them.

There is an enormous variety of different kinds of structures within living organisms. They occur at all levels of size, from an elephant's trunk to organelles within a cell, visible only with the electron microscope. There is still no satisfactory classification of the great range of processes by which these structures are brought into being. The following paragraphs constitute a tentative categorization that seems appropriate for the present state of biological thought on this topic.

Morphogenesis by differential growth. After their initiation, the various organs and regions of an organism may increase in size at different rates. Such processes of differential growth will change the overall shape of the body in which they occur. Processes of this kind take place very commonly in animals, particularly in the later stages of development. They are of major importance in the morphogenesis of plants, where the overall shape of the plant, the shape of individual leaves, and so on, depends primarily on the rates of growth of such component elements as the stems, the lateral shoots, and the vein and intervein material in leaves. In both animals and plants, such growth processes are greatly influenced by a variety of hormones. It is probable that factors internal to individual cells also always play a role.

Although differential growth may produce striking alterations in the general shape of organisms, these effects should probably be considered as somewhat superficial, since they only modify a basic pattern laid down by other processes. In a plant, for instance, the fundamental pattern is determined by the arrangement of the lateral buds around the central growing stem; whether these buds then grow fast or slowly relative to the stem is a secondary matter, however striking its results may be.

Morphogenetic fields. Many fundamental processes of pattern formation (*e.g.,* the arrangement of lateral buds in growing plants) occur within areas or three-dimensional masses of tissue that show no obvious indications of where the various elements in the pattern will arise until they actually appear. Such masses of tissue, in which a pattern appears, have been spoken of as "fields." This word was originally used in the early years of the 20th century by German authors who suggested an analogy between biological morphogenetic fields and such physical entities as magnetic or electromagnetic fields. The biological field is a description, but not an explanation, of the way in which the developing system behaves. The system develops as though each cell or subunit within it possessed "positional information" that specifies its location within the field and a set of instructions that lays down the developmental behaviour appropriate to each position.

The appearance of pattern

There have been several attempts to account for the nature of the positional information and of the corresponding instructions. The oldest and best known of these is the gradient hypothesis. In many fields there is some region that is in some way "dominant," so that the field appears as though organized around it. It is suggested that this region has a high concentration of some substance or activity, which falls off in a graded way throughout the rest of the field. The main deficiency of the hypothesis is that no one has yet succeeded in identifying satisfactorily the variables distributed in the gradients. Attempts to suppose that they are gradients of metabolic activity have, on investigation, always run into difficulties that can only be solved by defining metabolic activity in terms that reduce the hypothesis to a circular one in which metabolic activity is defined as that which is distributed in the gradient.

Recently, a new suggestion has been advanced concerning position information. Most processes within cells normally involve negative feedback control systems. These systems have a tendency to oscillate, or fluctuate regularly. In fact, any aspect of cell metabolism may be basically oscillatory in character; the cycle of cell growth and division may be only one example of a much more widespread phenomenon. The substances involved in these oscillations are likely to include diffusible molecules capable of influencing the behaviour of nearby cells. It is easy to envisage the possibility that there might be localized regions with oscillations of higher frequency or greater amplitude that act as centres from which trains of waves are radiated in all directions. It has been suggested that positional information is specified in terms of differences in phase between two or more such trains of transmitted oscillations.

Certain types of field phenomenon may involve an amplification of stochastic (random) variations. In systems containing a number of substances, with certain suitable rates of reaction and diffusion, chance variation on either side of an initial condition of equilibrium may become amplified both in amplitude and in the area involved. In this way, the processes may give rise to a pattern of differentiated areas, distributed in arrangements that depend on the boundary conditions.

Morphogenesis by the self-assembly of units. Complex structures may arise from the interaction between units that have characteristics such that they can fit together in a certain way. This is particularly appropriate for morphogenesis at the simple level of molecules or cells. Units such as the atoms of carbon, hydrogen, oxygen, nitrogen, and so on, can assemble themselves into orderly molecular struc-

Molecular assembly

tures, and larger molecules, such as those of tropocollagen, or protein subunits in general, can assemble themselves into complexes whose structure is dependent on localized and directional intermolecular forces. It seems that such comparatively large entities as the units that come together to form the head structures of bacteriophages or bacterial flagella are capable of orderly self-assembly, but the chemical forces that give rise to the interunit bonds are still little understood.

Processes that fall into the same general category as self-assembly may occur within aggregates of cells. The units that self-assemble are the cells themselves. Interaction and aggregation may be allowed to occur in assemblages of cells of one or more different kinds. In such cases it is commonly found that the originally isolated cells tend to adhere to one another, at first more or less at random and independently of their character, but later they become rearranged into a number of regions consisting of cells of a single kind. When the cells in the initial collection differ in two different characteristics, for instance in species and organ of origin, the assortment in some cases brings together cells from the same organ, in other cases cells from the same species. Mixtures of chick and mouse cells, for instance, reassort themselves into groups derived from the same organ, whereas cells from two different species of amphibia sort out into groups from the same species more or less independently of organ type.

This morphogenetic process probably has only a restricted application to the formation of structures in normal development, in which only in a few tissues (*e.g.*, the connective system) do cells ever pass through a free stage in which they are not in intimate contact with other cells, and cells of different origin do not normally become intermingled so as to call for processes of reassortment. To explain normal morphogenetic processes of plants and animals one must look to the results that can be produced by the differential behaviour of cells that remain in constant close contact with one another. Several authors have shown how striking morphogenetic changes could be produced within a mass of cells that remain in contact, but that undergo changes in the intensity of adhesion between neighbouring cells, in the area of surface in the proportion to cell volume, and so on.

Differentiation. Differentiation is simply the process of becoming different. If, in connection with biological development, morphogenesis is set aside as a component for separate consideration, there are two distinct types of differentiation. In the first type, a part of a developing system will change in character as time passes; for instance, a part of the mesoderm, starting as embryonic cells with little internal features, gradually develops striated myofilaments, and with a lapse of time develops into a fully formed muscle fibre. In the second type, space rather than time is involved; for instance, other cells within the same mass of embryonic mesoderm may start to lay down an external matrix around them and eventually develop into cartilage. In development, differentiation in time involves the production of the characteristic features of the adult tissues, and is referred to as histogenesis. Differentiation in space involves an initially similar (homogeneous) mass of tissue becoming separated into different regions and is referred to as regionalization.

Tissues and regions

Histogenesis involves the synthesis of a number of new protein species according to an appropriate timetable. The most easily characterized are those proteins formed in a relatively late stage of histogenesis, such as myosin and actin in muscle cells. The synthesis of proteins is under the control of genes, and the problem of histogenesis essentially reduces to that of the genetic mechanisms that direct protein synthesis.

Regionalization is concerned with the appearance of differences between various parts of what is at first a homogeneous, or nearly homogeneous, mass. It is a prelude to histogenesis, which then proceeds in various directions in the different regions so demarcated. The processes by which the different regions acquire distinct contrasting characteristics must be related to some of the processes discussed under morphogenesis. Unlike morphogenesis, regionalization need not involve any change in the overall spatial shape of the tissues undergoing it. Regionalization falls rather into the type of process for which field theories have been invoked.

Control and integration of development

PHENOMENOLOGICAL ASPECTS

One of the most striking characteristics of all developmental systems is a tendency to produce a normal end result in spite of injuries or abnormalities that may have affected the system in earlier stages. In many cases, perhaps in most, only injuries inflicted during a certain restricted period of development can be fully compensated for. During such periods the system is said to be capable of regulation or the restoration of normality.

Developmental regulation is often discussed in terms of homeostasis, or regulatory mechanisms. Many systems, including biological ones, exhibit a tendency to return to initial equilibrium once they are diverted from it. A developing system is, by definition, always changing in time, moving along some defined time trajectory, from an initial stage, such as a fertilized egg, through various larval stages to adulthood, and finally to senescence. The regulation that occurs in such systems is a regulation not back to an initial stable equilibrium, as in homeostasis, but to some future stretch of the time trajectory. The appropriate word to describe this process is homeorhesis, which means the restoration of a flow.

A second major phenomenological characteristic of development is that the end state attained is not unitary but can be analyzed into a number of different organs and tissues. The overall time trajectory of this system can, therefore, also be analyzed into a number of component trajectories, each leading to one or another of the end products that can be distinguished in the later stages. A major discovery of the early experiments on developing systems was that, in many cases at least, the different time trajectories diverge from one another relatively suddenly during some short period of development, which usually occurs well before any visible signs of divergence can be seen microscopically or by any other available means of analysis. The most dramatic and influential example of this was provided by studies on the development of the amphibian egg at the time of gastrulation, or formation of a hollow ball of cells. At this time the lower hemisphere of the embryo will be pushed inward (invaginated) to develop into the mesoderm and endoderm, and the upper hemisphere will remain on the surface, expanding in area to cover the whole embryo. Approximately one-third of the upper hemisphere will develop into the nervous system and the remainder into the skin. During the period when these morphogenetic movements of invagination and expansion are occurring, a process takes place by which a portion of the upper hemisphere enters a trajectory toward neural tissue and another part enters a trajectory leading to epidermal development. This process of determination of developmental pathways happens relatively quickly, during a period when the cells of the two different regions appear superficially alike. The occurrence of the determination can in fact be demonstrated only experimentally. Before it occurs, any part of the hemisphere can develop either into neural tissure or into skin. After it has happened, each part can develop only into one or the other of these alternatives.

Determination

It is clear that an adequate theory of development has to account not only for the processes by which a developing system moves along its appropriate time trajectory, but also for the nature of the processes by which the trajectories diverge from one another and become fixed or determined in the developing cells.

The determined state can be transmitted through many cell generations. An example of this transmission can be seen in *Drosophila* flies. The imaginal buds of *Drosophila* are small packets of cells that become separated from the main body of the embryo in the early stages of development. They persist throughout larval life and then enter into the differentiation of adult characteristics when stimulated to do so by the hormones secreted at the time of pupation. These pupation hormones disappear from the

body of the adult insect, and imaginal buds transplanted into the body cavity of an adult undergo many cell generations, but they do not show any signs of differentiating into the specific tissues of the corresponding adult organ. After many generations of proliferation, however, the cells can be transplanted back into a larva ready to pupate; they thus submit to the pupation hormones and differentiation occurs. Through many generations of proliferation the cells have retained the determination as to which adult organ they will develop into when the pupation hormones become available.

Attempts to identify the determining agent have not yet been successful. Experiments on amphibian eggs, however, have given rise to one important general conclusion; namely, that the process of determination can take place only during a certain period of development, in which the cells of the upper half of the amphibian egg are poised between the two alternatives of development into neural tissue or into skin. They are said at this time to be "competent" for one or the other of these types of development. While they are in this state, and only while they are in it, a variety of external agents can switch them into one or the other of the possible pathways. Such a situation may be contrasted with one in which the cells were neutral, or featureless, and required then an external agent to transmit to them the quality of becoming nervous tissue or of becoming skin. This would mean that the reacting cells required information or instructions to be added to them from outside. Such a situation is not characteristic of biological development. Both in highly developed organisms such as amphibians and in simpler ones such as bacteria, the external agents act only as a releaser that switches on one or another process for which all of the necessary information is already incorporated in the cells concerned.

ANALYTICAL ASPECTS

The existence of these developmental phenomena was realized in the first third of this century. During this period, biologists had no clear notion of the fundamental concepts needed to explain development. Developmental biologists, or embryologists, attempted to account for their observations by means of ill-defined notions, such as "potencies" or "organ-forming substances," or by referring to cellular properties that are real enough but obviously in themselves complex and essentially secondary in nature, such as cellular adhesiveness, the capacity of cell surfaces to differentially absorb certain substances, and so on. It was only gradually that developmental biologists came to realize the importance of the demonstration by genetics that nearly all the instructions required for the building of a new organism are contained in the genes that come together during fertilization, and that the small additional amount of information, contained primarily in the ovum, is itself a product of genetic instructions provided in the body of the mother in which the ovum is produced. The fundamental problems of the theory of development are, therefore, to understand how these units interact with one another to form more complex mechanisms that bring about the cellular or tissue behaviours of the different types of developing systems.

The genes as repositories of developmental instructions

In the development of the neural system of vertebrates, for example, a great many genes must be active in controlling the synthesis of particular proteins. In the formation of the wing of a *Drosophila,* the activity of some 20 or 30 genes has been definitely demonstrated, and certainly many more are involved. The action of all these genes, however, must be considered to form a network involving many types of feedback and other interactive loops, the overall result of which is a product in which many components are present in precisely defined concentrations; and further, the developmental process leading to this end result must be buffered or stabilized, in the sense that if the process is diverted from its normal course at an early stage, it returns to some later stage of the normal trajectory. The realization that the basic units of development are genes indicates that a stabilized time trajectory involves the action of tens, if not hundreds, of genes. The realization that biological development is fundamentally an expression of the controlled activities of genes has finally resolved one of the old philosophical controversies about the nature of development, between preformation and epigenesis. The former supposed that, at the initiation of development, for instance in the fertilized egg, the system already contained some representative of every organ that would eventually put in an appearance. The vindicated theory of epigenesis, on the other hand, supposed that later appearing entities were produced during the course of development.

The modern interpretation of epigenesis is that the initial stage of development does contain certain entities with well-defined properties, namely the genes. These do not, however, represent directly the later formed organs, which arise by the gradual interaction and progressive unfolding of the properties of groups of genes.

One of the major problems confronting modern developmental biology—namely, the nature of "determination"—requires an understanding of how genes are "primed" to enter into activity when an appropriate stimulus is given. The state of priming presumably has to apply to quite a large number of genes, though perhaps not to all that will be involved in the stabilized, or buffered, time trajectory, since some may be brought into activity by the operation of the earlier active ones. The priming, moreover, has to be able to persist through cell division and be capable of transmission through many generations of cell proliferation. Few concrete suggestions as to the mechanism have yet been made. One is that the primed genes are already producing the ribonucleic acid molecules, called messenger RNA's, which direct protein synthesis in the cell, but that these messengers are in some way inactivated or prevented from activating the protein-synthesizing machinery; this is known as the "masked messenger" hypothesis. Arguments in favour of this hypothesis are, however, circumstantial rather than direct. In some cases, for instance that of the *Drosophila* imaginal buds, there is direct evidence against it. Another hypothesis, perhaps more attractive, but much vaguer, is that the determination or priming involves the intervention of some of the large amounts of reiterated DNA known to be present in the cells of higher organisms. At the present time, however, biology lacks any convincing theory of determination in terms of gene action.

The "masked messenger" hypothesis

It appears at first sight that more is known about actual differentiation than initial determination. Actual differentiation must involve the controlled synthesis of particular proteins, coded for by specific genes. Certainly, a great deal is known about the mechanisms that control the action of genes in directing the synthesis of proteins in simple organisms such as viruses and bacteria. It is tempting to suppose that similar systems operate in controlling the synthetic activities of genes in higher organisms. Unfortunately, no single case of an exactly similar controlling system has ever been discovered in higher organisms, in spite of an intense search for it. It may in fact be suggested that until there is a fuller understanding of the mechanism of "priming" genes at the time of determination, there can scarcely be an adequate account of the way in which the activity of these genes is controlled at later stages.

Development and evolution

Evolution is carried out by a process dependent on mutation and natural selection. Expositions of this thesis, however, tend to overlook the fact that mutation occurs in the genotype, whereas natural selection acts only on the phenotype, the organism produced. It follows from this that the theory of evolution requires as one of its essential parts a consideration of the developmental or epigenetic processes by which the genotype becomes translated into the phenotype. The consequences of such considerations are discussed in the following sections.

EFFECT ON LIFE HISTORIES

Length and timing of the reproductive phase. Natural selection results in the production by one generation of offspring that are able to survive and reproduce themselves to form a further generation. The time unit appropriate to natural selection is therefore the generation interval. There will always be some natural selective pressure for the shortening of the generation interval, simply out of

a natural economy, and for an increase of the number of offspring produced by any reproducing individual. One of the ways in which such an increase could be assured would be the lengthening of the reproductive phase in the life history; another would be an increase in the number of offspring produced.

These are, of course, not the only natural selective pressures that operate. It is clear enough that, in evolution, they have often been overcome by other pressures. There is another natural selective pressure of more general importance. This is the pressure to restrict the length of the reproductive period, and indeed to remove reproductive individuals, in order to make room for the maturation of a new generation in which new genetic combinations can be tried out for their fitness. A species whose individuals were immortal would exhaust its possibilities for future evolution as soon as its numbers saturated all the ecological niches suitable for its way of life. Death is a necessary condition for the trying out of new genetic combinations in later generations. It is usually brought about, in great part at least, by combinations of two processes: restriction of the period of effective reproduction to a certain portion of a life history, and as a necessary consequence of this, the absence of natural selection for genetic mutations that would be effective in preserving life after reproduction has ceased. In some organisms—for instance, long-lived trees—there may be no restriction of reproduction to a particular period in the life history, but their development involves the gradual accumulation of larger and larger quantities of nonliving materials, such as dead wood, which presents a growing handicap, in the face of which the organism cannot indefinitely maintain itself against the inevitable hazards of existence. It is still something of a question whether these natural selective forces are sufficient in themselves to account for the phenomena of senescence, aging, and eventual death, which are found in various forms throughout nearly the whole biological kingdom (see below *Aging and senescence*).

The evolutionary significance of death

As was mentioned above, evolution has produced a number of the types of multiphasic development, in which the life history involves a succession of larval stages. Such types of development offer the possibility of changing the relative importance of the various stages in relation to the exploitation of resources and reproduction by the species. There are, for instance, many types of animals (particularly insects) in which nearly the whole life history is passed in a larval stage in which most of the feeding and growth of the organism is carried out, the final adult stage being short and used almost entirely for reproduction. Another evolutionary strategy has been to transfer the reproductive phase from the final stage of the life history to some earlier larval stage. This again has occurred in certain insects. If such a process is carried to its logical evolutionary conclusion, the final previously adult stage of the life history may totally disappear, the larval stage of the earlier evolutionary form becoming the adult stage of the later derivative of it. An example in which this process is at least partially accomplished is in the axolotl, a salamander that reproduces in a larval stage and in nature rarely if ever metamorphoses into the adult, but can be persuaded to do so if injected with extra supplies of the hormone thyroxin. It has been suggested that such processes of neoteny (the retention of some juvenile characteristics in adulthood) have played a decisive role in certain earlier phases of evolution, evidence of which is now lost. It has been argued that the whole vertebrate phylum may have originated from modifications of one of the larval stages of an invertebrate group.

The role of neoteny in evolution

Recapitulation of ancestral stages. The modifications of life histories just mentioned are aspects of a more general situation; namely, that the only variations that can become available for natural selection to operate on are those that can be produced by alterations of the developmental or epigenetic system of an existing organism. Any new mutant gene can cause a change only in a preexisting set of developmental interactions; the phenotypes to which it can give rise are limited by the nature of the system that it will modify. One immediate result of this situation is that the development of a later evolved form will retain many features from the development of its ancestors: most evolutionary developments are likely to be additions to the previous organization. Since there is evolutionary pressure to reduce the length of time between generations, the addition of a new feature to development is likely to be accompanied by a speeding up of the older stages, and probably omission of certain of them.

To repeat, the development of a late-evolved form retains those aspects of earlier life histories that are essential for the building up of later developmental stages that may be important for natural selection. In the vertebrates, for instance, highly evolved types such as mammals and birds produce during their early development remnants of the primitive kidneys (pronephros and mesonephros) that functioned as excretory organs in their evolutionary ancestors. Although these organs no longer perform their physiological functions in later organisms, they play an essential role during the formative processes of embryonic development. Some structures characteristic of evolutionary ancestors may be retained for relatively short evolutionary periods after they have lost their original function simply because there is not sufficient natural selective pressure to bring about their elimination when they no longer have any obvious function, either physiologically or epigenetically; the human appendix is an example.

Adaptability and the canalization of development. A developing organism is subjected to natural selection by its particular environment. The environment is not the same for all individuals of a population, nor does it necessarily remain the same throughout evolutionary periods of time. An organism can be regarded as having to meet environmental changes that are unpredictable. There are basically two different types of strategy employed, in various proportions in different organisms, to meet this situation. One, perhaps the more obvious, is to evolve a high capacity for modification by environmental circumstances in ways that increase fitness in the environment in question; this is the strategy of increasing adaptability. It is probably true to say that all organisms show some capacity for adaptation, either short-term (physiological) or longer term (developmental), to their environments. In most organisms, however, particularly in most higher organisms, there is considerable development of the alternative strategy, which is to build up well-buffered or channelled developmental processes, which lead to the production of a relatively predictable invariant end result in the face of very diverse environments. The second strategy is likely to be followed in situations in which the environment is likely to change markedly during the course of the organism's life.

The tendency to channel development

Whether or not this is the main reason for the evolution of channelled, or canalized, developmental systems, a considerable degree of canalization is very common. It is relatively rare to find instances in which the form of an animal is highly dependent on the early environment, although such dependence is common enough among plants. Much more frequently, situations such as that typified by the house mouse are encountered: the mouse develops into an almost identical form whether it lives in the tropics or in a cold-storage depot.

This canalization of development severely restricts the phenotypic effects that can be produced by mutations. In particular, many new mutations occurring in a single dose in a diploid organism are found to be recessive, or ineffective in causing any alteration in the phenotype. As this discussion makes clear, canalization should not be considered as a relation involving only the normal and mutated forms of a particular gene, but rather the result of the interaction of many genes.

GENETIC ASSIMILATION

A long-standing controversy in biology has been concerned with whether phenotypic modifications produced by abnormal environments are heritable in the sense that they can be produced by later generations in the absence of the original environmental stress. The hypothesis that they are heritable was advanced by the French evolutionist Lamarck in the 18th century and is generally known as the "inheritance of acquired characters." It found some

supporters among biologists, some of whom used it as an argument against the Darwinian theory of evolution. In a broad sense, all characters are to some extent inherited, in that they depend on the genotype of the organism, and to some extent acquired, since development is also affected by the environment. In a stricter sense, however, Lamarck's hypothesis suggests that there is some inherent biological property that enables organisms to pass on physical modifications to their descendants, independently of a Darwinian mechanism of selection.

The combination of adaptability and canalization in development can explain such phenomena in strictly Darwinian rather than Lamarckian terms. The abnormal environment acting during development may succeed in modifying even a well-canalized development system. If the modification is of an adaptive kind and increases the fitness of the individuals in the unusual environment, it will be favoured by natural selection. The development of the selected individuals will, however, also show some properties of canalization, that is to say, resistance to further environmental changes. This invariance may be sufficient to prevent offspring of the selected individuals from reverting completely to the original phenotype even if they are removed from the abnormal environment. After selection for an adaptive modification in an abnormal environment has proceeded for many generations, a form may be produced whose canalization is strong enough to maintain the new phenotype almost unaltered when the environment reverts to what it was before the abnormality occurred. This process, which has been demonstrated in a number of laboratory experiments, is known as genetic assimilation. It produces exactly the same results as those emphasized by advocates of the Lamarckian inheritance of acquired characters, but it produces them by an orthodox Darwinian mechanism operating on developmental systems that have the common properties of canalization and adaptability. It provides the most convincing explanation for the evolution of organisms that are physiologically or functionally adapted to the demands their way of life will make. (C.H.W.)

PLANT DEVELOPMENT

Although both plants and animals share the chemical basis of inheritance and of translation of the genetic code into structural units called proteins, plant development differs from that of animals in several important ways. Higher plants sustain growth throughout life and, in this sense, are perpetually embryonic; animals, on the other hand, generally have a determinate period of growth, after which they are considered mature. Furthermore, both growth and organ formation in plants are influenced by their possession of a rigid cell wall and a fluid-filled space called the vacuole, two features unique to the plant cell. Conversely, certain features of animal cells are absent in plants. Notable is the lack of cellular movements and fusions that play an important part in tissue and organ development in higher animals.

General features

LIFE CYCLES

Development of gametes

The life cycle of all tracheophytes (vascular plants), bryophytes (mosses and liverworts), and many algae and fungi is based on an alternation of generations, or different life phases: the gametophyte, which produces gametes, or sex cells, alternating with the sporophyte, which produces spores. Gametophytes develop from the spores and, like them, are normally haploid; *i.e.*, each cell has one set of chromosomes. Sporophytes develop from a fertilized egg, or zygote, that results from the fusion of gametes (fertilization) formed by the gametophytes; they are accordingly diploid; *i.e.*, each cell has two sets of chromosomes. Although the two generations are phases of one life cycle, they have independent developmental histories; each begins as a single cell, passes through a juvenile period, matures, and gives rise to the alternate phase.

In various algae and fungi the two generations are alike in form (*i.e.*, are isomorphic), and, despite the difference in chromosome number, their development follows essentially identical pathways. More commonly, however, the alternating generations have different forms (*i.e.*, are heteromorphic); this is true for the bryophytes and for all vascular plants, both angiosperms (flowering plants) and gymnosperms (conifers and allies). General rules for vascular plants are that the sporophyte generation is physically the larger, has a more complex developmental history, produces a greater range of cell types, and expresses a more diverse biochemistry; the gametophyte is often diminutive, reduced in the case of the angiosperms to a mere few cells. In the bryophytes, the gametophyte generation, rather than the sporophyte, is the more conspicuous.

Although the gametophyte generation in vascular plants is small and has limited physiological capabilities, its cells must convey genes capable of directing the sporophytic developmental pattern, because the pattern is transmitted through the gametes to the zygote (Figure 6). The expression of "sporophytic" genes must therefore be repressed in the gametophyte, probably from the time of spore formation (sporogenesis). Correspondingly, events associated with gamete formation (gametogenesis) or fertilization must somehow free the sporophytic genes and thus permit the zygote to enter the sporophytic developmental pattern. Although it might be supposed that the "switch" is associated with the difference in chromosome number between the haploid spore (a single set) and the diploid zygote (a double set), this has been shown not to be the determining factor.

Conveying the genes through alternate generations

The alternation of generations illustrates an important principle, namely that cell lineages arising from single parental cells containing the same genetic potentiality may pursue mutually exclusive developmental patterns. Channelling, or canalizing, events of this nature occur repeatedly in the course of development of an individual plant, beginning with the pattern of cell division from the very first cleavage of the zygote cell.

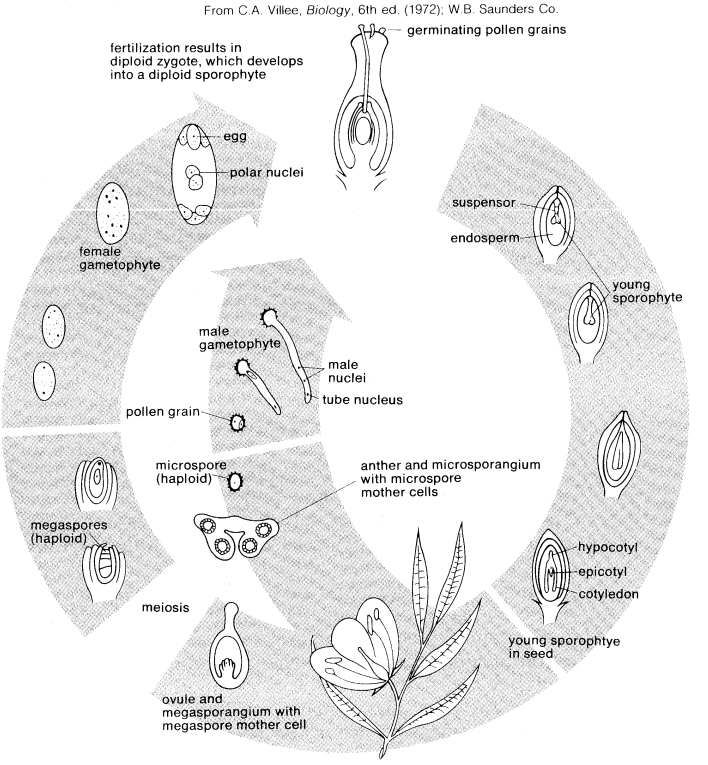


Figure 6: Stages in the fertilization and seed formation of a dicot.

BODY PLANS

Collectively plants manifest a wide range of body plans, ranging from the single cell (or unicell), with a single nucleus, through various types of colonial and filamentous forms to massive multicellular structures. (Algae, including the single-celled forms, have a great deal in common in structure and biochemistry with vascular plants. Bacteria and fungi generally are not considered plants, but because of their various plantlike qualities they are taken as plants for the purposes of this section.)

For the unicell, development is the same as cell differentiation. Although many unicellular fungi and algae show little differentiation other than that connected with reproduction, others undergo elaborate structural changes that illustrate many principles basic to development in multicellular plants. An important example is the green alga *Acetabularia.* This alga first produces a rootlike system and stalk and then, later, a flattened umbrella-like cap. The developmental potentialities of this unicell, with its single nucleus, are, however, limited; in order for there to be any advance beyond the state seen in *Acetabularia,* with the development of greater body mass and a division of labour among different parts, an increase in the number of participating nuclei seems obligatory.

One method of providing more nuclei is by nuclear division without a corresponding cell division; the result is a coenocytic structure. Plants with this type of multinucleate organization show considerable diversity; examples are found in both algae and fungi. Growth occurs by the extension of the cell wall in certain zones, usually at the tips of filaments, and structural differentiation results from branching and the specialization of parts for particular functions. The aggregation of coenocytic filaments can lead to the development of a three-dimensional body, or thallus, but plants with this type of organization have not achieved great size.

A more significant type of body plan, one based on the multicellular filament, is found in its most simple form in certain algae known as diatoms, in which chains of cells of indefinite length arise, although the cells show no evidence of interaction. More advanced is the condition of many other algae, in which there occur branches that may either be identical with the original filament or show structural or physiological specialization. This condition occurs in certain green algae, in which the main branches creep and the lateral branches grow erect; such diversification represents an important developmental innovation and, possibly, the evolutionary beginnings of organ specialization in plants.

Multicellularity: an advance in development

Three-dimensional body forms may evolve from the association of cells in colonies. Cells among the colonial green algae are of definite number; each component cell resembles a free-living unicell, but all are united by cellular connections, or plasmodesmata, which may be important in coordinating the development of the colony. Colonies are often of precise geometric shape, forming either a circular plate or a sphere. In elaborate ones, certain cells are specialized for reproduction, and others are concerned primarily with movement.

Another developmental pathway resulting in more massive body structure is by the association of filaments. The reproductive structures of many fungi are composed of large numbers of closely interwoven filaments, which, although not physically connected, do interact in some way to produce structures such as the mushroom cap. Several filamentous body plans are found among the red algae. In one, a single main (axial) filament grows at one end, but, just behind its tip, cells divide to produce a number of lateral filaments that grow parallel with the axial filament. The older parts of the thallus, therefore, seem to be an aggregate of filaments. More massive structures are produced when there are several axial filaments; and, by branching, particularly when accompanied by fusion, dense tissues resembling the basic undifferentiated tissue (parenchyma) of higher plants are formed. In these algae, cellular connections occur between daughter cells of a filament, and others may develop secondarily between cells of neighbouring filaments.

The transition from a filamentous to a three-dimensional form appears most notably in the brown algae. In certain brown algae, growth is by an axial filament, but, behind the tipmost cell, divisions produce a denser tissue lacking evidence of filamentous organization. In the sporophytes of kelp, one of the largest and most complex of the algae, cell division often is restricted to areas comparable to the growing tips of vascular plants, and, although a filamentous organization may be evident in the centre of the thallus, the surrounding cortical regions are composed of a tissue that is essentially undifferentiated. (The gametophytes of kelp, however, have a simple filamentous organization.)

Three-dimensional body form

Among nonvascular plants, true parenchyma is found in the bryophytes, in both the gametophyte and sporophyte phases. The development of the moss gametophyte illustrates the transition from a filamentous to a highly organized three-dimensional growth form. The moss spore germinates into a filamentous plant, the protonema, which later produces a leafy shoot. This type of transition from simple to more complex growth form is accompanied by the synthesis of new kinds of ribonucleic acids (RNA's), presumably through the activation of genes that were not expressed during the early growth of the gametophyte.

Much of the remainder of this section is concerned with the development of the complex body forms of vascular-plant sporophytes, which do not normally pass through any filamentous stages. It may be noted, however, that, in the course of evolution, the capacity for this type of growth has not been lost, since it may be adopted by cells grown in tissue cultures in the laboratory.

PREPARATORY EVENTS

The sporophytes of all vascular plants produce cells called spore mother cells—since they will give rise to spores—in spore cases (sporangia). Spore mother cells are usually surrounded, during development, by a special nutritive tissue. In the more primitive groups each sporangium holds many mother cells. This is true also in the pollen-producing sporangia of gymnosperms and angiosperms but not in the egg-producing sporangia (ovules), which usually have only one mother cell.

In certain lower vascular plants, typified by the club moss *Selaginella,* the gametophyte is formed entirely—or almost entirely—within the spore wall. Two kinds of gametophytes develop from the two kinds of spores produced by the sporophyte in different sporangia; the larger spore (megaspore) gives rise to the female gametophyte, the smaller spore (microspore) to the male. This condition is referred to as heterospory. The gametophytes, or prothalli, of other club mosses and most horsetails and ferns are sexually undifferentiated and arise from one kind of spore, a condition termed homospory.

In these groups the gametophytes develop as free-living and independent plants that ultimately produce the gametes. In general, the male gametes (antherozoids) are produced in globose structures (antheridia) that are either stalked or sunken in the gametophyte. The antherozoids, always many in number, develop from mother cells enclosed in the jacket of the antheridium. Each antherozoid can move by using its whiplike hairs, or flagella, two or three (in the lycopods) or many (in the horsetails and ferns). The female gametes are formed singly in flask-shaped structures (archegonia) that also are either stalked or sunken in the gametophyte. The neck of the flask is closed by neck canal cells, which later break down to permit the entry of the male gamete. The egg itself lies in the basal part, or venter, of the flask, with a ventral canal cell above it. When the male gametes, or antherozoids, are released by the rupture of the antheridium, they swim in a water film to the archegonia and effect fertilization.

Among the gymnosperms the male gametophyte is much reduced and is a parasite on the sporophyte for only a short time. Cell cleavages within the spore wall cut off a prothallial cell, which will give rise to the vegetative (*i.e.,* nonreproductive) part of the plant, and an antheridial cell, which divides into a tube cell and a generative cell. The male gametophyte so formed and contained within the spore wall is the pollen grain. After transfer to the ovule by wind, the pollen grain germinates to form a tube, and

the generative cell divides into two cells, one of which forms the male gametes by further division. The gametes bear numerous spirally arranged flagella. The female gametophyte meanwhile develops entirely within the parent sporangium in the ovule. The size of the single functional spore increases greatly as the spore nucleus divides repeatedly to produce numerous free nuclei. Cell-wall formation then begins at the periphery, extending inward until the whole area is divided into cells. Up to four archegonia are formed, sunken in the tissue of the gametophyte, each with a female gamete, or egg.

Fertilization: a change of generations

The end of the gametophyte phase and the beginning of the sporophyte phase occur at fertilization, when one of the male gametes fuses with the female gamete to form the zygote, which will then develop as the sporophyte. (Development of the sporophyte can, in some cases, be triggered by means other than fertilization, in which case the organism is said to arise parthenogenetically.)

The male gametophyte of angiosperms is reduced to three cells, one so-called vegetative cell and two male gametes. The division producing the gametes may occur either before dispersal of the pollen grain or later, during the growth of the pollen tube. The female sporangium has one or two coats, or integuments, except for an opening (micropyle) at one end; the sporangium with an integument is called the ovule. The female gametophyte, known in this group as the embryo sac, develops from the parent spore while it is still retained in the sporangium. Three cell divisions result in eight nuclei, which arrange themselves so that three lie at each end and two lie in the centre. The cytoplasm then cleaves and three cells are formed at each pole, leaving two nuclei in a large central cell. The three cells at the micropylar pole (end toward the micropyle) form the egg apparatus. Two of these cells, called synergids, correspond to the neck cells of an archegonium; the third is the egg cell. The three cells at the opposite pole, the antipodals, play a part in embryo nutrition in certain genera. The two polar nuclei in the central cell ultimately unite, becoming the fusion nucleus. The pollen grain is transferred by various agencies (wind, water, animals) to the stigma of the female flower, and, as in the gymnosperms, it germinates to produce a tube. This tube grows through intervening tissues, through an opening (micropyle) of the egg, and enters a cell near the micropyle (synergid), in which the two male gametes are discharged. The unique feature of this phase of angiosperm development is that two fertilizations occur. One male gamete fuses with the egg to give the diploid zygote; the other makes its way to the fusion nucleus in the central cell, already diploid, and by a second fusion gives a triploid primary endosperm nucleus, which is later concerned in the formation of the nutritive tissue, or endosperm.

Early development: from zygote to seedling

EMBRYO FORMATION

Cleavage of the zygote. In vascular plants embryo formation, or embryogenesis, usually occurs within a few hours after fertilization, with the first cell division that cleaves the zygote, or fertilized egg, into two daughter cells. Thereafter, rapid cell division provides the building blocks of the primary organs of the embryo sporophyte: the first root, first leaves, and the shoot apex. Temporary structures concerned with embryo nutrition—suspensor and foot—may also be produced. These organs originate in a polarization established at the time of zygote cleavage, but the details of their development vary widely among the different groups.

In the club mosses the zygote divides in a plane at right angles to the axis of the archegonium. The daughter cell toward the neck forms a short filament of cells, the suspensor; the inner cell gives rise to the other organs of the embryo, the shoot, root, and foot. The axis of the embryo is inclined to that of the archegonium and may be almost at right angles. This is in contrast to the behaviour of the true mosses, in which the embryo is oriented along the length of the archegonium, with the foot directed inward and the structures that are equivalent to the shoot, namely the spore capsule and its stalk, directed toward the neck.

A polarity like that of the mosses appears in the horsetails, in which the zygote divides by transverse and longitudinal walls to form a group of four cells. Of these, the two cells toward the neck give rise to the shoot system; the inner two produce the foot and root.

Embryogenesis in gymnosperms and angiosperms

The details of early embryogenesis in gymnosperms vary consider-ably. In the cycads and ginkgos, the initial cleavage establishes a polarity opposite to that in the horsetails, the inner cell giving rise to the shoot and the outer producing the root. Many conifers are unique in that the zygote undergoes a period of free-nuclear division without cell formation, producing usually four or eight nuclei, which move to the end of the zygote, away from the neck cells, where cleavage begins. In the pines a further division gives four tiers of four cells. The intermediate tiers extend greatly to form a suspensor; each of the four cells at the lower pole may act as the parent cell of an embryo, a condition sometimes referred to as polyembryony.

In contrast, there is no free-nuclear stage in angiosperm embryogenesis. The zygote cleaves by a wall more or less at right angles to the axis of the embryo sac. The daughter cell next to the micropyle (basal cell) produces a suspensor and contributes to the root; the inner (terminal) cell gives rise to the shoot system. (Angiosperm embryogenesis is more fully described in the following section dealing with the origin of primary organs.)

Notwithstanding the variation in the different groups, the pattern of development established in the early cell cleavages is consistent. The primary polarization of the zygote must necessarily be imposed by the adjacent tissues of the sporophyte, but thereafter the fate of daughter cells depends on control established within the young sporophyte itself.

Although it is often possible to specify the origin of the cell lineages contributing to the various organs and tissue layers, a geometric regularity in cell division is generally maintained through only the first few division cycles in the embryo. The final form of the embryo is thus determined not through the specification of a precise scheme of cell division, as in the development of colonial algae, but through an overall control in which cell and tissue interactions play an important part.

Origin of the primary organs. Angiosperm embryogenesis can be described in terms of a much studied flowering plant called shepherd's purse (*Capsella bursa-pastoris*). The zygote divides into two cells, the terminal cell and the basal cell. The terminal cell divides by a wall formed at right angles to the first cleavage wall and then again by a wall formed at right angles to this; a quadrant of cells is thus formed (Figure 7A–D). The partition of the quadrant cells in a transverse plane then produces an octant stage (Figure 7E). By transverse divisions, the basal cell forms a filament, the suspensor, of up to ten cells, the end cell of which swells to form an absorbing organ. The attachment cell, or hypophysis, adjoins the octants derived from the terminal cell (Figure 7F).

At this time, the prospective future of each of the zones of the embryo can be specified. Four cells of the octant group will ultimately produce the seed leaves (cotyledons) and the shoot apex; the other four will form the hypocotyl, the part of the embryo between the cotyledons and the primary root (radicle). The hypophysis will give rise to the radicle and the root cap; the cells of the suspensor will degenerate as the embryo matures.

Establishing the embryo

The zones of the embryo destined to form the principal organs are established by this first sequence of divisions, and tissue layers are defined during the ensuing divisions. The octant cells divide by curved walls parallel to the surface; in this way the outer layer responsible for producing the epidermis of the shoot system is defined. Divisions of a more irregular nature in the inner zone ultimately define the tissues from which the central cylinder and vascular core of the main axis of the shoot will develop. Simultaneously, the hypophysis forms a group of eight cells by three successive divisions, the planes of which are mutually at right angles. Of these eight cells, the outer four produce the root cap and epidermis; the inner four contribute to the radicle.

The embryo is at first globular (Figure 7F), but it soon

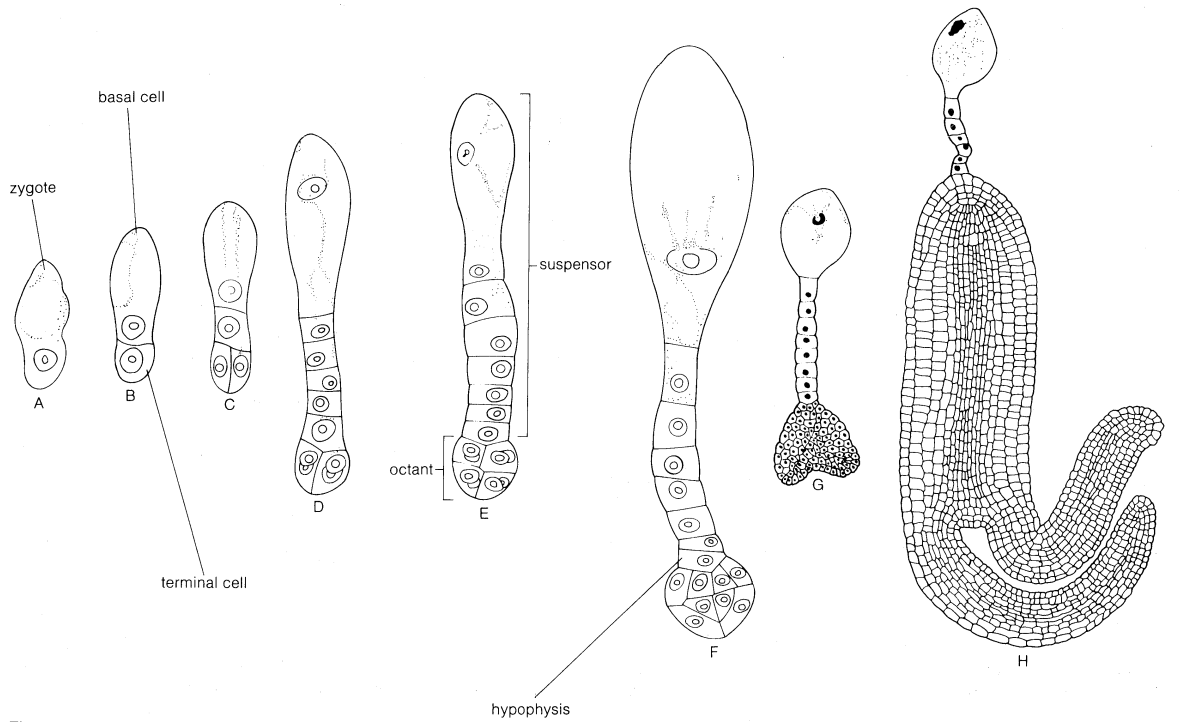


Figure 7: Development of the embryo in shepherd's purse *Capsella bursa-pastoris* (see text; G and H less enlarged than A–F).

After Soueges and Schaffner in P. Maheshwari, *Introduction to the Embryology of the Angiosperm*, copyright © 1950; used with permission of McGraw-Hill Book Company

becomes heart-shaped by a combination of numerous cell divisions and enlargement in two zones of the outer hemisphere (Figure 7G). In this manner two cotyledons form. The volume of tissue between the cotyledons is the prospective shoot apex. The characteristic form of the apex is not established until after germination.

As the cotyledons become extended, the embryo bends, because of physical restraints, to conform with the cavity of the embryo sac (Figure 7H). From the heart-shaped phase onward, the core of the hypocotyl and the radicle appears as a cylinder of narrow and elongated cells. This is the parent tissue of the vascular system of the seedling. The surrounding tissue contributes the cortex layer of the stem and root.

The embryogenesis of *Capsella* illustrates only one of several patterns found among flowering plants. Among dicotyledons, the planes of division of the terminal cell, the form of the suspensor, and the contribution made by the basal cell to the embryo all provide evidence used in determining the embryogenetic plan.

Monocotyledons, flowering plants the seeds of which contain only one cotyledon, share with dicotyledons such as *Capsella* the main features of early embryogenesis, including the possession of a suspensor and, in most cases, a fairly regular progression of cell divisions to the octant stage. Thereafter the symmetrical growth pattern is lost through the development of the single cotyledon. In the lily family (Liliaceae), generally accepted as a primitive family of monocotyledons, the cotyledon is derived from an octad of cells arising from the terminal cell. The hypocotyl and stem apex are derived from the proximal cell of a short filament formed by the basal cell, and the root comes from the pair of cells next to it. The suspensor forms from the distal cell or cells of the filament. In the more advanced families of monocotyledons, including the grasses (Gramineae) and orchids (Orchidaceae), embryogenesis is much less regular. The grass embryo possesses structures that do not occur in any other flowering plants, namely, the scutellum, an organ concerned with the nutrition of the seedling, and the coleoptile and coleorhiza, protective sheaths of the young shoot and the radicle. The scutellum arises from octant cells, which also contribute to the cotyledon. The basal cell forms part of the coleoptile and also gives rise to the shoot apex and the tissues of the root and coleorhiza. The embryo is asymmetrical, with the shoot apex lying on one side in a notch, ensheathed by the coleoptile.

In marked contrast, embryogenesis of the orchids is more simple. Except when a suspensor is formed, early cleavages follow no well-defined plan, and the product is an ovoid mass of tissue called the proembryo. No cotyledon, stem apex, or root apex is organized in this early period; these organs do not appear until after germination has occurred.

Embryos as parasites

Nutritional dependence of the embryo. During their early growth, the embryos of all vascular plants exist as virtual parasites depending for nutrition on either the gametophyte or the previous sporophyte generation through the agency of the gametophyte or, in the special case of the angiosperms, upon an initially triploid tissue, the endosperm, which is itself nourished by the parent sporophyte.

The early nutrition of the sporophyte in ferns, horsetails, and club mosses such as *Lycopodium* is clearly provided by the gametophyte. In these groups the young sporophyte produces a multicellular structure, the foot, which remains embedded in the tissues of the gametophyte throughout early development withdrawing nutrients. Ultimately, both shoot and root of the sporophyte grow out from the gametophyte, but, even after the first leaf has begun to photosynthesize and thus to produce its own food, the gametophyte may persist.

In *Selaginella,* the gametophytes are sexually distinct. The female gametophyte develops within the wall of the megaspore. The archegonia are exposed after the megaspore wall splits, but the gametophyte never escapes completely. After fertilization, the zygote cleaves, and the outer cell produces a long suspensor that pushes the embryo deeply into the tissues of the gametophyte. A foot is then formed, as in *Lycopodium,* and further development of the embryo continues at the expense of reserves transferred to the megaspore from the preceding sporophyte generation.

There are superficial similarities between the nutritional history of the embryo in gymnosperms and in *Selaginella,* for, in each, the female gametophyte, dependent upon reserves derived from the sporophyte, acts as an intermediary between one sporophyte generation and the next.

In the pines, the female gametophyte develops within the tissues of the nucellus and acquires abundant food reserves. The proembryo forms after a period of free-nuclear division in the zygote, and the tier of cells above the basal

four then elongates to form a suspensor, which pushes the embryonic group deep into the gametophyte (Figure 8). Secondary suspensor cells may form from the basal tier to continue the process. During embryogenesis, the gametophyte continues to grow and to accumulate food materials, which are transferred to the embryo or remain as reserves in the seed.

From *Comparative Morphology of Vascular Plants* by Adriance S. Foster and Ernest M. Gifford, Jr. W.H. Freeman and Company. Copyright © 1959

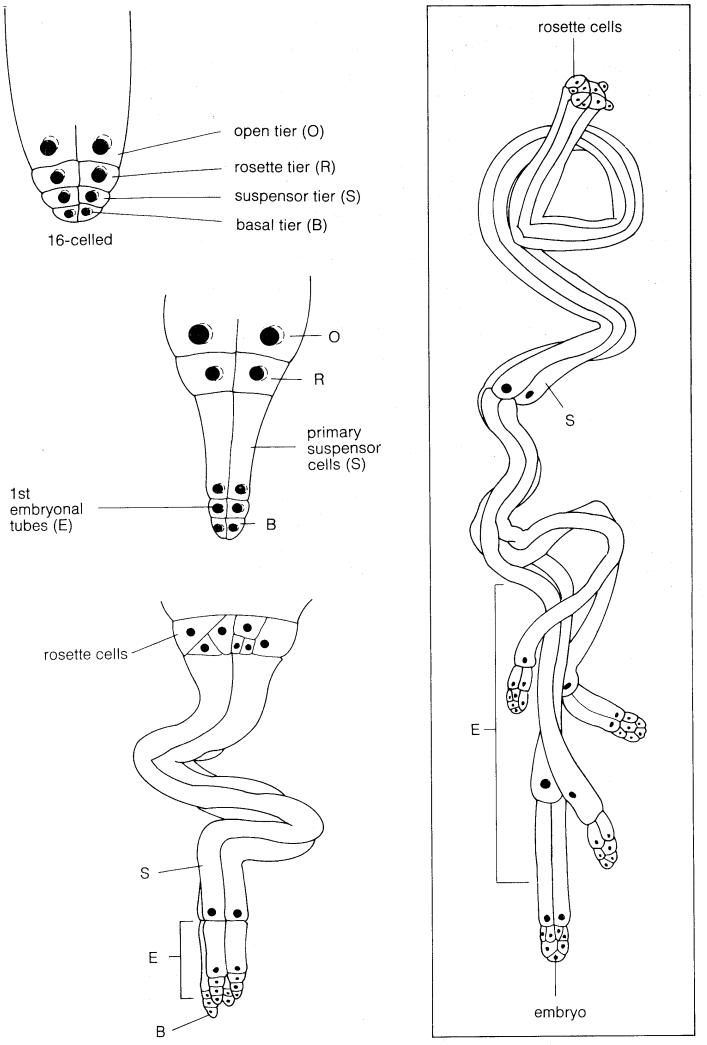


Figure 8: Fertilization and early embryogenesis in *Pinus*.

The female gametophyte of angiosperms never acquires copious reserves, although starch is frequently present in the central cell and sometimes in the egg itself. The unique feature, here, is that the embryo is nutritionally dependent upon the endosperm, a tissue that, in the genetical sense, constitutes a third organism—neither gametophyte nor sporophyte. Furthermore, as a tissue the endosperm manifests several other special characteristics. The nuclei have three chromosome sets and, therefore, three times the deoxyribonucleic acid (DNA) of haploid cells. As nuclear division ends, the amount of DNA per nucleus increases still further, a condition comparable with that in various plant- and animal-gland nuclei, presumably connected with the nutritional function of the endosperm. Nuclear division takes place at first without cell-wall formation so that a coenocyte is produced; later, partitioning of the cytoplasm results in a cellular tissue.

The food reserves

The reserves accumulated in the endosperm include carbohydrates (especially starch), lipids, and proteins. As reserves accumulate, the nuclei of the endosperm cells may undergo deformation and degeneration. In many plants the growing embryo consumes the endosperm before seed maturation; in others, the tissue persists in the seed, providing a reserve for the developing seedling after germination. Endosperm is not formed in certain angiosperms. In such cases the embryo depends on the transfer of nutrients directly from the sporophyte.

Tissues other than the endosperm may become specialized for the early nutrition of the embryo. The antipodal cells of the female gametophyte sometimes acquire glandular properties, as may cells of the nucellus surrounding the embryo sac. In some species the embryo itself develops a suspensor that penetrates the tissues of the parent sporophyte and acts as an absorbing organ.

Dormancy of the embryo. Among the lower pteropsids (club mosses, horsetails, and ferns), the principal agent of dispersal is the haploid spore and not, as in gymnosperms and angiosperms, the seed, the ripened ovule containing a dormant embryo. Since the embryo of lower pteropsids is not involved in dispersal, it does not usually undergo any marked period of dormancy after the differentiation of the primary organs. Development instead proceeds continuously through dependence upon the gametophyte until the young sporophyte is established as a physiologically independent plant. The embryos of gymnosperms and angiosperms pass into a state of dormancy soon after the differentiation of the primary organs and the sporophyte is dispersed in a seed.

In the period leading up to dormancy, several changes occur in the embryo. The accumulation of reserves in the cotyledons or elsewhere ceases, respiratory rate declines rapidly, and cell division, with associated protein and nucleic-acid synthesis, stops. Correlated with these events are cellular changes typical of tissues with low metabolic activity. Especially obvious is the general dehydration of the cells that constitute the seed and the thickening of the cell walls of the ovule to form the seed coat (testa). The product is a structure in which the embryo is protected from temperature extremes by its state of desiccation and is often guarded from further drying and from mechanical or biological degradation by the seed coats. The seed coat often contributes to the maintenance of dormancy by physically impeding the passage of water and gases to and from the embryo, by chemically inhibiting germination, and by mechanically restricting the growth of the embryo.

Role of the seed coat in dormancy

GERMINATION AND EARLY GROWTH

Dormancy is brief for some seeds, for example those of certain short-lived annual plants. After dispersal and under appropriate environmental conditions, such as suitable temperature and access to water and oxygen, the seed germinates, and the embryo resumes growth.

The "breaking" of dormancy. The seeds of many species do not germinate immediately after exposure to conditions generally favourable for plant growth but require a "breaking" of dormancy, which may be associated with change in the seed coats or with the state of the embryo itself. Commonly the embryo has no innate dormancy and will develop after the seed coat is removed or sufficiently damaged to allow water to enter. Germination in such cases depends upon rotting or abrasion of the seed coat in the soil. Inhibitors of germination must be either leached away by water or the tissues containing them destroyed before germination can occur. Mechanical restriction of the growth of the embryo is common only in species that have thick, tough seed coats. Germination then depends upon weakening of the coat by abrasion or decomposition.

In many seeds the embryo cannot germinate even under suitable conditions until a certain period of time has lapsed. The time may be required for continued embryonic development in the seed or for some necessary finishing process—"after ripening"—the nature of which remains obscure.

Low temperatures and seed germination

The seeds of many plants that endure cold winters will not germinate unless they experience a period of low temperature, usually somewhat above freezing. Otherwise germination fails or is much delayed, with the early growth of the seedling often abnormal. (This response of seeds to chilling has a parallel in the temperature control of dormancy in buds.) In some species, germination is promoted by exposure to light of appropriate wavelengths; in others, light inhibits germination. For the seeds of certain plants, germination is promoted by red light and inhibited by light of longer wavelength, in the "far red" range of the

spectrum. The precise significance of this response is as yet unknown, but it may be a means of adjusting germination time to the season of the year, or of detecting the depth of the seed in the soil. Light sensitivity and temperature requirements often interact, the light requirement being entirely lost at certain temperatures.

In the process of germination, water is absorbed by the embryo, which results in the rehydration and expansion of the cells. Shortly after the beginning of water uptake, or imbibition, the rate of respiration increases, and various metabolic processes, suspended or much reduced during dormancy, resume. These events are associated with structural changes in the organelles (membranous bodies concerned with metabolism), in the cells of the embryo.

The emergence of the seedling. Active growth in the embryo, other than swelling resulting from imbibition, usually begins with the emergence of the primary root from the seed, although in some species (*e.g.,* the coconut) the shoot emerges first. Early growth is dependent mainly upon cell expansion, but, within a short time, cell division begins in the radicle and young shoot; thereafter, growth and further organ formation (organogenesis) are based upon the usual combination of increase in cell number and enlargement of individual cells.

Until it becomes nutritionally self-supporting, the seedling depends upon reserves provided by the parent sporophyte. In angiosperms these reserves are found in the endosperm, residual tissues of the ovule, or in the body of the embryo, usually in the cotyledons. In gymnosperms, food materials are contained mainly in the female gametophyte. Since reserve materials are partly in insoluble form—as starch grains, protein granules, lipid droplets, and the like—much of the early metabolism of the seedling is concerned with mobilizing these materials and delivering, or translocating, the products to active areas. Reserves outside the embryo are digested by enzymes secreted by the embryo and, in some instances, also by special cells of the endosperm.

In some seeds (*e.g.,* castor beans) absorption of nutrients from reserves is through the cotyledons, which later expand in the light to become the first organs active in photosynthesis. When the reserves are stored in the cotyledons themselves, these organs may shrink after germination and die or develop chlorophyll and become photosynthetic.

Orientation of the seedling

Environmental factors play an important part not only in determining the orientation of the seedling during its establishment as a rooted plant but also in controlling some aspects of its development. The response of the seedling to gravity is important. The radicle, which normally grows downward into the soil, is said to be positively geotropic. The young shoot, or plumule, is said to be negatively geotropic, because it moves away from the soil; it rises by the extension of either the hypocotyl, the region between the radicle and the cotyledons, or the epicotyl, the segment above the level of the cotyledons. If the hypocotyl is extended, the cotyledons are carried out of the soil, but, if the epicotyl elongates, the cotyledons remain in the soil.

Light affects both the orientation of the seedling and its form. When a seed germinates below the soil surface, the plumule may emerge bent over, thus protecting its delicate tip, only to straighten out when exposed to light (the curvature is retained if the shoot emerges into darkness). Correspondingly, the young leaves of the plumule in such plants as the bean do not expand and become green except after exposure to light. These adaptative responses are known to be governed by reactions in which the light-sensitive pigment phytochrome plays a part. In most seedlings, the shoot shows a strong attraction to light, or a positive phototropism, which is most evident when the source of light is from one direction. Combined with the response to gravity, this positive phototropism maximizes the likelihood that the aerial parts of the plant will reach the environment most favourable for photosynthesis.

Later development: the sporophyte plant body

CONTINUATION OF ORGAN FORMATION

Although it is convenient to refer to the early development of the plant sporophyte from the fertilized egg as embryogenesis, the process is never actually concluded as it is in the higher animals. In vascular plants, organ formation (organogenesis) is not confined to early life, and the processes of shoot, root, and leaf formation that occur first in the embryo are repeated, albeit in modified form, throughout the life of the plant. The life-span may be short and determinate, as in annual plants such as the cereals, or long, lasting for many years—indeed potentially indefinitely, except for limitations imposed by the environment and accidents—as in trees. The protracted growth of perennials, or plants that resume growth each growing season, tends to lead to increase in size, but bulk is not necessarily directly correlated with age, because individual leaves, flowers, and even whole limbs continuously die and are shed. Some long-lived plants, however, do reach a point at which losses of body mass balance the increase resulting from continued growth and organ formation.

The activity of meristems. Characteristically, vascular plants grow and develop through the activity of organ-forming regions, the growing points. The mechanical support and additional conductive pathways needed by increased bulk are provided by the enlargement of the older parts of the shoot and root axes. New cells are added through the activity of special tissues called meristems, the cells of which are small, intensely active metabolically, and densely packed with organelles and membranes, but usually lacking the fluid-filled sacs called vacuoles. Meristems may be classified according to their location in the plant and their special functions. One important distinction is between persistent meristems, typified by those of the growing points, and meristems with a limited life, those associated with organs, such as the leaf, of determinate growth. The regions of rapid cell division at the tips (apices) of the stem and the root are terminal meristems. In the stem apex, the uppermost part is the promeristem, below which is a zone of transversely oriented early cell walls, the file, or rib, meristem. The procambium is a meristematic tissue concerned with providing the primary tissues of the vascular system; the cambium proper is the continuous cylinder of meristematic cells responsible for producing the new vascular tissues in mature stems and roots. The cork cambium, or phellogen, produces the protective outer layers of the bark.

Classification of meristems

Among meristems of limited existence is the marginal, or plate, meristem responsible for the increase in surface area of a leaf; it contributes new cells mainly in one plane. Another type of meristem of limited life is called intercalary; it is responsible for the extension of some stems (as in the grasses) by the addition of new tissues remote from the growing points.

The number of dividing cells in persistent meristems remains roughly constant, with one of the daughter cells of each division remaining meristematic and the other differentiating as a component of a developing organ. The geometrical arrangements in the particular organ determine the way in which this occurs, but in general the consequence is that the meristem is continuously moving away from the maturing tissue as growth continues. It remains, therefore, a localized zone of specialized tissue, never becoming diluted by the interposition of expanding or differentiating cells. In organs such as leaves, flowers, and fruits, in which the growth is determinate, the divisions of meristematic cells become more widely scattered, and the frequency progressively falls as the proportion of the daughter cells that differentiate increases. Ultimately, at maturity, no localized meristem remains.

The contribution of cells and tissues. The two major factors determining the forms of plant tissues and organs are the orientation of the planes of cell division and the shapes assumed by the cells as they enlarge. Clearly, if the division planes in a cell mass are randomly oriented and individual cells expand uniformly, the tissue will enlarge as a sphere. On the other hand, if cell division planes are oriented regularly or the expansion of individual cells is directional, the tissue can assume any of a number of shapes. In a stem, for example, the cell division planes of the promeristem are oriented at various angles to the stem axis, so that new cells produced contribute to both width and length. Below this region, in the rib meristem, the proportion of divisions with the cell plate at right angles to

the axis increases, so that the cells tend to be oriented in files. The cells in these files expand vertically more than they do horizontally, and, accordingly, the stem develops as a cylinder.

Influence of the mitotic spindle

The factors that control the orientation of cell division planes in meristems are largely unknown. Cell interactions, however, are presumed to coordinate the distribution and orientation of the divisions. In each cell microtubules in the cytoplasm help to orient the nucleus before it divides. Then, at the time of the division, other microtubules arranged in a spindle-shaped figure (the mitotic spindle) are involved in separating the daughter chromosomes and moving them to opposite ends of the parent cell. Thereafter, the residual part of the spindle helps to locate the plate that separates the two daughter cells. Microtubules are also concerned in determining the direction of growth in expanding cells, since they appear to influence the construction of the cell wall by controlling the way cellulose is laid down in it.

Although change in shape is a form of cell differentiation, the term in the more general sense refers to a change in function, usually accompanied by specialization and the loss of the capacity for further division. Biochemical differentiation often involves a change in the character of the cell organelles—as when a generalized potential pigment body (proplastid) matures as a chloroplast, a chlorophyll-containing plastid. But it may also involve structural changes at a subcellular level, as when organelles change their character in cells engaged in intense metabolic activity.

Plant cell walls

The differentiation of plant cells for the movement of materials and the provision of mechanical support or protection invariably depends upon modification of the walls. This usually entails the accretion of new kinds of wall materials, such as lignin in woody tissue and cutin and suberin in epidermal tissues and cork. The accompanying structural changes must be controlled, for the wall materials are not applied at random but according to a pattern appropriate to the particular cell or tissue. The development of patterns during cell-wall growth depends not only on the cytoplasmic microtubules, as in the construction of the cells that will give rise to the water-conducting vessels (xylem elements), but also on cytoplasmic membranes, as in the formation of sievelike end walls (sieve plates) in the cells that will give rise to food-conducting vessels (phloem elements).

The differentiation of xylem culminates in the death of the participating cells, and the vessels are formed of chains of empty walls. This is an example of "programmed death," not an uncommon phenomenon in plant and animal development.

THE SHOOT SYSTEM AND ITS DERIVATIVES

The shoot tip. The gametophytes of mosses and liverworts and the sporophytes of many higher plants have a shoot, or early stem, with a single cell at its tip, or apex, from which all the tissues of the stem arise. This apical cell is usually four-sided (tetrahedral), with three faces directed downward, and the fourth capping the apex. Daughter cells are continually cut off sequentially from the three inner faces, the apical cell preserving its tetrahedral shape. In cell lineages derived from the daughter cells, the division planes may remain oriented in a more or less regular manner, so that, for some distance below the apex, the three sectors can be recognized in the stem. This basic pattern occurs in the arrangement of the "leaves" of some mosses, which lie in three ranks. In many plants, however, division planes in the lower part of the apex show no particular correlation with the planes of cleavage of the apical cell, and the lateral appendages do not reflect any three-part arrangement.

The generative role of promeristem tissue

Gymnosperm and angiosperm apices do not possess apical cells. The generative role is discharged by an ill-defined zone of tissue called the promeristem. Regularities may appear in the distribution of division planes only in the extreme tip region. Over the outer part of the apex, the cells often appear to lie in one to three layers, which constitute the tunica (Figure 9). Enclosed by the tunica lies a core of cells that exhibits no distinct layering; this zone is

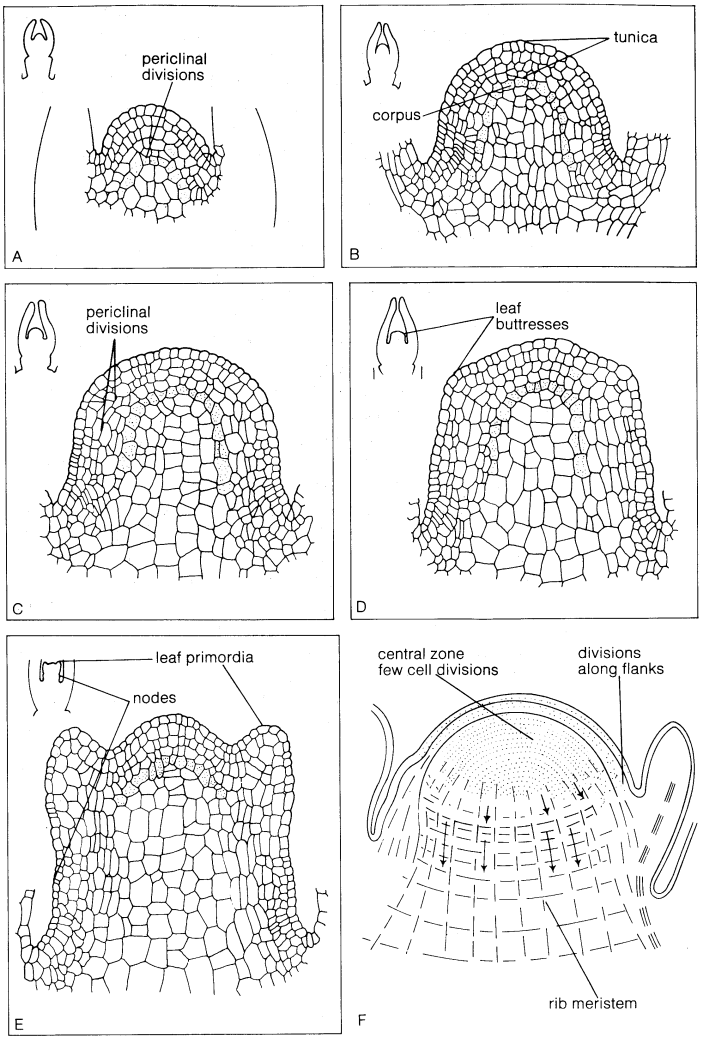


Figure 9: Leaf initiation in shoot tip of *Hypericum uralum*.
From (A–E) Zimmerman in K. Esau, *Plant Anatomy* (1965); John Wiley & Sons, Inc.; (F) F.C. Steward and M. Ram, *Advances in Morphogenesis*, 1:245, Academic Press, Inc.

the corpus. The layers of the tunica normally contribute to the surface layers of the plant, and the corpus provides the deeper lying tissues.

The tunica–corpus analysis emphasizes the orientation of division planes, but apices can be examined from other points of view—the sizes of cells, the degree of vacuolation, and the concentration of various cell constituents, especially ribonucleic acid (RNA), vary through the apex and this sometimes results in more or less distinctive zones. Both gymnosperm and angiosperm apices have been classified on the bases of such zonal patterns, but the validity of this approach, as well as its usefulness for understanding the function of the apex as a morphogenetic centre, has been questioned.

Since 1950, a theory of angiosperm apical zonation developed by French and Belgian botanists has been gaining support. This theory proposes that the central region of the apical dome constitutes a mass of cells with relatively low division rates, the *méristème d'attente,* or "waiting meristem." Surrounding this region is an annular zone of cells with higher division rates, the *anneau initial,* or "initiating ring." Features other than division rates characterize these zones: RNA and protein content are lower in the *méristème d'attente* than in the *anneau initial,* and the nucleoli are smaller. In longitudinal section, the differences contribute to the patterns distinguishable in apices, some of which have been used as bases for structural classification. The main contention of the Franco-Belgian school, however, is that the zonation represents a functional difference. The *méristème d'attente* is regarded as a region mainly concerned with controlling the geometry of the apex. The cells have a restricted metabolism concerned primarily in

maintaining a low rate of increase in cell number, and they themselves, as well as their immediate derivatives, take no part in organogenesis or associated differentiation. The *anneau initial,* by contrast, is that part of the apex that produces the beginnings, or primordia, of lateral organs. Not only is the division rate higher, but the tissue as a whole is involved in metabolic syntheses that precede morphogenesis.

One difficulty in investigating the stem apex arises from the uncertainty about which aspects are important for the overall function: division planes, division frequency, metabolic patterns, or some combination of these. Still another complication results because the apex is in a state of constant change during the growth of the plant. A long-term developmental trend begins after the definition of the growing point in early embryogenesis and continues thereafter through juvenility and the period of vegetative growth into the reproductive phase. Superimposed on this trend is a cyclical change reflecting the periodic generation of the primordia of leaves and lateral shoots in the region immediately under the apex.

The production of leaves. Leaves originate on the flanks of the shoot apex. A local concentration of cell divisions marks the very beginning of a leaf; these cells then enlarge so as to form a nipple-shaped structure called the leaf buttress. The cells of the leaf buttress may be derived from the tunica alone or from both the tunica and the corpus.

In the early growth of the leaf primordium, new cells are contributed mainly by meristematic activity at the pole directed away from the stem, so that the buttress extends in length. The subsequent distribution of growth varies among the different groups of vascular plants according to the shape of the mature leaf. In considering the angiosperms, a broad-leaved dicotyledon (tobacco; Figure 10) and a narrow-leaved monocotyledon (maize [corn]) will serve as examples.

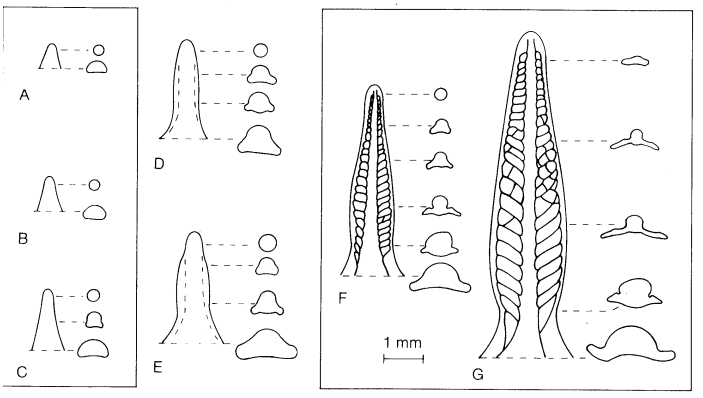


Figure 10: *Longitudinal and transverse sections of tobacco foliage leaf primordia in successive stages of development.* (A,B) The primordium in early development. (C) The small lateral ridges appearing about midway on the primordium are forerunners of thelamina. (D) Development of the lamina started. (E) Further expansion of the lamina and the beginning of development of the main lateral veins. (F) Primordium five millimetres in length showing development, toward the base, of the network of provascular strands. (G) Primordium with the development of the lamina and system of venation well started. Dotted lines in C–E represent external boundaries of midrib and lateral veins and not the provascular portion.

The tobacco leaf as an example

Apical growth dominates in the tobacco-leaf primordium until a height of about 0.5 millimetre (0.02 inch) is reached. Thereafter, the buttress becomes more and more flattened in the transverse plane by laterally oriented cell divisions and further expansion growth on either side. The dividing zones are the marginal meristems, through the activity of which the leaf gains its laminate form. In each meristem the outer file of cells, or marginal initials, contributes the epidermal layers by continued division. The cells below, the submarginal initials, provide the tissue of the inner part of the leaf. Usually a certain number of cell layers is defined in the mesophyll (the parenchyma between the epidermal layers of a foliage leaf). Cell division is not limited to the region of the marginal meristems but continues throughout the leaf in each of the layers, always in the same plane, until the final cell number is approached. The rate then declines, ceasing in the different layers at different times. Divisions usually end first in the epidermis, then in the lower mesophyll layers of a leaf such as that of tobacco, and last in the main photosynthetic tissue, the palisade layer, just beneath the upper epidermis.

The vascular pattern in a tobacco leaf is determined early in the development of the vessel primordium. A procambial strand is formed by the elongation of narrow axial cells, and this extends both toward the base and toward the apex, eventually linking with the procambium of the stem. When the marginal meristems become active, the lateral veins of the leaf are initiated first, followed by the third and later order branchings that give the characteristic network of veins in the mature leaf.

Although the differentiation of the cells of the vascular system begins at the base, the epidermal and mesophyll cells mature from the tip inward toward the stem. The palisade cells elongate in a plane at right angles to the epidermis; those of the lower mesophyll expand irregularly to give lobed forms. The cells of the epidermis, shaped like irregular paving stones, continue to expand in the plane of the leaf after growth ceases in the mesophyll, so that the cells of the internal tissues are pulled apart to form the system of air spaces found in the mature leaf.

Dicotyledonous leaves are folded in various ways in the bud, the patterns being determined by differential growth in the tissues of the upper and lower surfaces (laminae) of the young leaves. Differential growth may cause the lamina either to roll or fold toward the leaf midrib or to fold near lateral veins, thus pleating the lamina. The folds in the bud are, of course, eliminated during the final phase of leaf expansion.

In the development of the maize leaf, the primordium arises first as a prominence some distance below the apical dome. The zone of division and growth extends laterally around the apex so that a complete collar forms; then the margins overlap. Meanwhile the original tip zone continues to elongate, eventually surpassing the stem apex. Tip growth declines thereafter, and further increase in cell number results from meristematic activity at the base. The early development of the vascular system is unlike that in dicotyledons, for several parallel procambial strands, rather than a single midrib, are initiated. The first of these grow toward the apex, but, as tip growth ceases, procambial strands form above and extend toward the base, passing through the node, or point of insertion of the leaf primordium, and into the stem below. As the leaf extends in length, the tissues begin to mature first at the tip, and a wave of differentiation passes down toward the base, where cell division and extension growth may continue long after the tip of the leaf is mature. Protection for this immature and succulent tissue of the leaf base is afforded by the sheaths of older leaves surrounding it.

Variations in leaves

These examples illustrate the principles involved in leaf development, but there are many deviations associated with variations in leaf form. Lobing and toothing result from the persistence of cell division and growth in particular stretches of the margin after growth ceases in between. Carried to the extreme, this localized growth gives the feather-like pinnate leaf. Many monocotyledons form cylindrical leaves as a result of a fusion of the margins of the primordium after it has encircled the stem.

Branching of the shoot. The shoots of most vascular plants branch according to a consistent plan, with each new axis arising in the angle between a leaf and a stem—that is, in a leaf axil. In some plants, buds may also form from the older parts of shoot or root remote from the main apices; these buds, termed adventitious, do not conform to the general plan.

A lateral shoot apex is initiated on the flanks of the main apex but at some distance below the point of emergence of the youngest leaf primordium. As in the origin of a leaf, generally the outer cell layers contribute to the surface tissues of the new apex by maintaining a consistent pattern of divisions. In some species a tunica of more than one cell layer quickly forms, so that the new apex appears as a miniature version of the main one; alternatively, the differentiation may not become apparent until the new

primordium has attained considerable bulk. In all cases, the new apex must reach a minimal volume before it in turn can begin to form its own lateral primordia and to organize true axillary buds. As this volume is attained, *méristème d'attente–anneau initial* zonation appears. As in the main apex, the formation of new primordia is associated with the annular zone.

From this point on, the development of the lateral shoot is the same as that of the main shoot, except that growth may not be as rapid because the main apex, or leading bud, dominates and absorbs much of the available nutrient. The early growth of the axillary bud proceeds quite vigorously until a certain number of leaf primordia has been formed; then apical activity slows. Cell division gradually stops, and with it the associated syntheses; thus there is no increase in the DNA of the nuclei of the meristem after the last division. The bud, in effect, passes into a state of dormancy, even though the external conditions for growth are propitious. This phenomenon is known as correlative bud inhibition, since it is determined by the activity of the leading bud of the shoot. If the leading bud is removed, the inhibited lateral buds resume growth, and with it the associated syntheses.

Vascular development. Cell division planes in the zone just below the apex of the shoot tend to be oriented so that vertical files of cells are formed. This is more evident in the central core than in the surrounding cortical region, for the pattern is not disturbed by the insertion of lateral members. The first signs of the differentiation of the vascular system appear some distance below the apex, in a zone of tissue distinguishable by the smaller cross-sectional area of individual cells. These cells, forming the procambium zone, arise by divisions oriented at right angles to the axis and may form a complete cylinder; generally, however, interruptions occur, the segments being related to the uppermost leaf primordia. In a dicotyledon such as tobacco, the cylinder at its highest level consists of strands running upward toward the points of insertion of the primordia. Thus, as the site of each primordium is determined, a strand forming in the adjacent region of the stem will contribute to the cylinder, but at a higher level than the preceding strand. The link with the earlier formed procambium is not simple, however. The strand passing upward toward a leaf primordium usually is composed of branches arising from strands that enter the two nearest older leaves below it. Because it in turn will contribute a branch for the next leaf, the cylinder is really a hollow network, the "gaps," or leaf traces, marking the points of departure of the leaf veins.

The "cylinder" of vessels

During subsequent development, the strands, or vascular bundles, increase in thickness by further cell divisions, and connections form with the vascular systems of axillary buds. The cells differentiate to give the characteristic tissues of the vascular system: phloem vessels (conducting tissue), phloem parenchyma (packing tissue), and phloem fibres (supporting tissue) toward the outside and xylem vessels (woody conducting tissue) toward the inside. The differentiation occurs in an upward direction, so that the maturation of the vascular tissues follows at a more or less constant distance behind the apex.

Although details differ, the above account of the origin of the primary vascular system is broadly applicable to gymnosperms and many ferns. Vascular development differs somewhat in certain flowering plants. In many monocotyledons, such as maize, the several vascular strands that pass down from each leaf primordium into the stem do not contribute to a single cylinder but are scattered in the ground tissue, or parenchyma, of the stem. Lateral interconnections form principally at the nodes.

Cambium activity

Increase in stem diameter is accomplished in the older stems of dicotyledons by the activity of the cambium, which produces secondary vascular tissue. This meristem, a relic of the procambium, is composed of thin-walled cells, is flattened in the radial plane, and persists between primary vascular tissue, the differentiated outer phloem, and the inner xylem. When secondary thickening begins, the parenchymatous cells between the vascular bundles also resume division, ultimately forming a cambium cylinder. The cells of the cambium divide, producing initial phloem cells toward the outside and initial xylem cells within. Files of cells are also cut off among the initial phloem and xylem cells that remain parenchymatous and are called phloem and xylem rays.

As in the apical meristem, the number of dividing cells in the cambium remains constant, except that more cells occasionally are added by divisions in the radial plane so that the girth of the cambial cylinder expands in pace with the growth of the xylem within. The addition of new phloem toward the outside compresses the primary phloem and the cortical tissues in a radial direction while stretching them tangentially. These primary tissues do not persist, however. As the girth of the stem increases, the epidermis is disrupted, and the outer layers of the cortex become meristematic, giving rise to the cork cambium, which generates cork cells on the outside. The cork layer, or bark, then takes over the protective function of the epidermis.

THE ROOT SYSTEM AND ITS DERIVATIVES

The root tip. Plants that have a single apical cell in the shoot also have a single apical cell in the root. The cell is again tetrahedral, but sometimes daughter cells are cut off from all four faces, with the face directed away from the axis producing the cells of the root cap. The cells derived from the other faces continue to divide mostly by forming transverse walls, but occasionally also in the longitudinal plane. In this way vertical columns of cells form—tending, because of their mode of origin, to be disposed in three sectors.

In the roots of gymnosperms, angiosperms, and some lower plants, there is no single apical cell. Again, as with the shoot, such root apices can be analyzed in different ways. Perhaps the most useful approach is based upon tracing the sources of the main tissues in the apical region. Such an analysis has led to the histogen theory, which proposes that the three principal tissues of the root—vascular cylinder, cortex, and epidermis—originate from three groups of initial cells, or histogens, in the apical meristem—plerome, periblem, and dermatogen respectively. A fourth histogen, the calyptrogen, produces the root cap. The histogens have been thought to lie in linear order in the apex, with the initial cells of the vascular system toward the older part of the root, and those of the cap toward the tip.

The histogen theory of root-tissue origin

The histogen theory is difficult to apply to some types of roots, and there has been uncertainty about the numbers of histogens. The discovery of the "quiescent centre" in the root apex has clarified many features, however. The quiescent centre is a group of cells, up to 1,000 in number, in the form of a hemisphere, with the flat face toward the root tip; it lies at the centre of the meristem, in much the same position, in fact, as the tetrahedral apical cell in certain lower plants (Figure 11). The cells of the quiescent centre are unusual in that their division rate is lower than that in the surrounding meristem. The cells of the centre have other distinctive features as well, notably a lower rate of protein synthesis than that of neighbouring cells.

The quiescent centre is surrounded by actively dividing cells of the promeristem that are the initial cells of the various tissues of the root. Those abutting the flat, tip-directed face contribute to the root cap; those above the quiescent centre are distributed in a cup shape. The cells in the centre of the cup produce the procambium and so, ultimately, give rise to the vascular cylinder. The annular zone of cells surrounding this central group forms the initials of the cortex; surrounding this, in turn, a ring of initial cells forms the protoderm, the layer corresponding to the epidermis at this level of the root.

The quiescent centre, a constant feature of the root tip, is apparently generally present in angiosperm and probably also in gymnosperms. The quiescent centre probably plays a role comparable with that of the apical cell in some lower plant roots, maintaining the geometry of the system. It has also been suggested that it may be concerned with the synthesis of growth hormones, although no direct evidence exists. When roots are damaged mechanically or by radiation, the cells of the centre can resume a rapid division rate, and they then participate in regeneration.

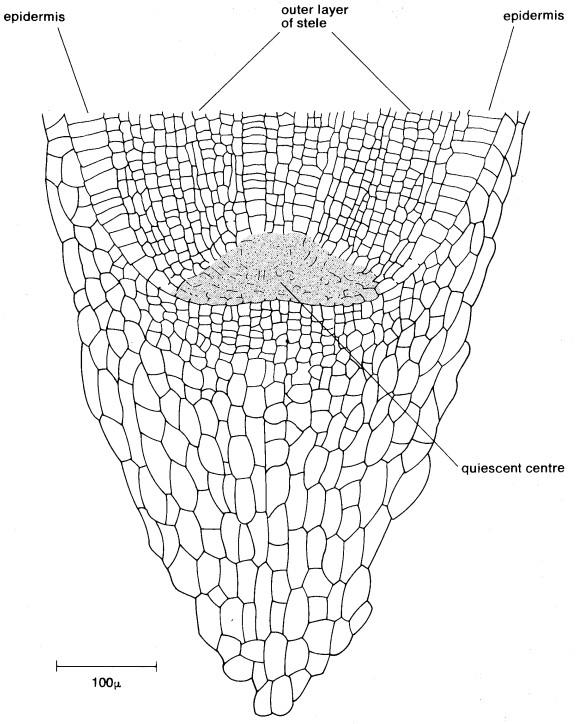


Figure 11: Median section of root apex of *Zea mays*.
Adapted from *Biological Reviews of the Cambridge Philosophical Society* (1959)

The zone of cell division extends some distance along the length of the root above the tip region. Although the girth may increase by longitudinal divisions and the widening of the daughter cells, most divisions occur in the transverse plane resulting in the formation of longitudinal files of cells.

In longitudinal section, the tissue zones become progressively better defined away from the tip. An internal protective band, the endodermis, becomes conspicuous as a single sheath of cells surrounding the procambium. The phloem procambium, recognizable by its narrow cells, begins to differentiate in the lower part of the region of elongation. The xylem also becomes distinct, the thickenings appearing first in the upper part of the extension zone. Differentiation keeps pace with the advance of the root tip as new cells are added in the promeristem. When xylem occupies the core, there is no pith as in the shoot, but the cells of the outermost layers of the vascular cylinder remain undifferentiated, forming the pericycle, a tissue important in the formation of lateral roots. Within the bounds of the pericycle, the xylem is star-shaped in section, with the first-formed xylem elements (protoxylem) occupying the ridges. The phloem lies in the intervening grooves. Outside of the endodermis, the cortical cells elongate but remain thin-walled. Above the level of the root-cap sheath, the epidermis forms the outer layer of the root, and, beyond the extension zone, its cells begin to develop root hairs.

A more complete account can be given for the mechanics of development of the root apex than for that of the stem, mainly because of its greater simplicity. An important difference lies in the absence of a mechanism for the cyclical production of lateral organs at the apex itself.

Genera-
ion of
ateral
oots

Branching of the root. The branching of the root takes place in the older parts and does not directly involve the apical meristem. The tissues concerned are the endodermis and the layer immediately beneath it, the pericycle. The endodermis participates in root branching in certain lower plants with apical cells. A cell of this layer enlarges and forms a tetrahedral cell, which becomes the new apical cell; by further divisions a hemispherical volume of tissue forms around it—the whole constituting a new apex.

In many other plants, including gymnosperms and angiosperms, the lateral roots develop from the pericycle. Cells in this layer enlarge and begin to divide until a dome of tissue develops. Called the incipient apex, the dome pushes out the surrounding endodermis, which may itself resume divisions, its daughter cells enlarging to create a sheath around the new root tip. During further growth, the dome assumes an organization like that of the primary root apex. At first, all cells are meristematic; then, while the primordium is still small, cells in the central zone cease DNA synthesis, and this zone becomes the new quiescent centre. Beyond it, the root cap is produced, and, at the base, initial cells begin to develop the cell files that become the vascular cylinder, cortex, and epidermis. The vascular tissues differentiate from the base outward, and link eventually with xylem and phloem of the parent root. All this development occurs before the tip of the new root emerges from the tissues of the parent root. The growth of the new tip into the cortex first pushes out the endodermal sheath, if one is present, and then bursts it. The cortical cells are themselves crushed and probably resorbed as the root grows on, until finally the tip breaks through the epidermis.

In most roots, new laterals are initiated in the pericycle opposite to the protoxylem ridges. They tend accordingly to form vertical ranks along the length of the root, reflecting the number of bands of protoxylem. Although lateral roots arise in quite a different way from leaves and axillary shoots at the stem apex, there are certain common features. Pericyclic cells about to produce a root primordium synthesize ribonucleic acid, in anticipation of the period of growth and morphogenesis that will result in a new apex. The same behaviour is seen in the cells of the annular zone, from which leaf primordia arise at the stem apex, and also in the axillary zones at a slightly lower level, from which new stem apices develop.

Later growth. In the secondary growth of the root, cell division in the primary xylem produces a cambium, which abuts the pericycle over the protoxylem ridges and passes between the phloem strands and the xylem in the grooves. Activity of the root cambium is comparable with that of the stem cambium; phloem elements are cut off outward, and xylem elements are cut off within. With continued growth in thickness, the star-shaped figure of the primary xylem is lost, and the cambium eventually forms a cylindrical sheath. Again, as in the stem, the protective function of the epidermis is ultimately taken over by cork layers produced by a cork cambium in the outer cortex.

Correlations in plant development

COORDINATION OF SHOOT AND ROOT DEVELOPMENT

Although the structural organization of the vascular plant is comparatively loose, development of the various parts is well coordinated. Control is dependent upon the movement of chemical substances, including both nutrients and hormones.

An example of correlation is the growth of shoot and root. The enlargement of aerial parts is accompanied by increased demands for water, minerals, and mechanical support that are met by coordinated growth of the root system. Several factors apparently are concerned with control, because shoot and root affect each other reciprocally. The root depends on the shoot for organic nutrients, just as the shoot depends on the root for water and inorganic nutrients and the flow of ordinary nutrients must, therefore, play some part. More specific control, however, may be provided by the supply of nutrients required in very small amounts. The root depends on the shoot for certain vitamins, and variation in the supply, reflecting the metabolic state of the aerial parts, may also influence root growth. In addition, hormonal factors affecting cell division pass upward from the root into the stem; although the exact role of the hormones has not yet been established with certainty, they may provide one way by which the root system can influence the activity of the shoot apex.

Meeting demands for food and support

The control of secondary thickening is another important example of growth correlation. As the size of the shoot system increases, the need for both greater mechanical support and increased transport of water, minerals, and manufactured food is met by an increase in stem girth through the activity of the vascular cambium. Generally, the cambium of trees in temperate zones is most active in

the spring, when buds open and shoots extend, creating a demand for nutrients. Cell division begins near the bud in each shoot and then spreads away from it. The terminal bud stimulates the cambium to divide rapidly through the action of two groups of plant hormones: auxins and gibberellins.

The inhibition of lateral buds, another example of correlated growth response, illustrates a reaction opposite to that occurring in the control of cambial activity. Lateral buds are inhibited in general because axillary shoots grow more slowly or not at all, while the terminal bud is active. This so-called apical dominance is responsible for the characteristic single trunk growth seen in many conifers and in herbaceous plants such as the hollyhock. Weaker dominance results in a bushy growth form with repeated branching. The fact that lateral, or axillary, buds become more active when the terminal bud is removed suggests that hormonal control is involved.

The flow of auxin from the shoot tip is, in part, responsible for inhibiting axillary buds. The nutritional status of the plant also plays a role, apical dominance being strongest when mineral supply and light are inadequate. Because axillary buds are released from inhibition when treated with cell-division promoting substances (cytokinins), it has been suggested that these substances are also concerned in regulating axillary-bud activity.

DETERMINATION OF MATURE FORM

The duration of a juvenile phase

After its establishment as an independent plant, the sporophyte passes through a juvenile period before reaching maturity and becoming reproductive. Juvenility may be brief or, as in the case of trees, may extend over several years. The duration is determined partly by internal factors and partly by environmental controls related to the seasons.

In some ways juvenility is a continuation of developmental trends initiated in the embryo. In many plants, new organs are produced sequentially through early life, each of progressively more mature form. The first leaf of the young fern sporophyte, for example, is small and relatively simple, and the vascular system consists of a few forked strands. As growth proceeds, succeeding new leaves are of increasing complexity, and the shape begins to resemble that typical of the reproductive frond; in addition, vasculation shifts to the mature pattern, often one with a network of veins. Comparable trends occur in flowering plants, in which leaves at successive levels of plant maturity often show a progressive increase in the complexity of lobing or toothing.

Some of the changes associated with the juvenile period can be attributed to the gradual enlargement of the growing point, necessarily small in the embryo; its volume increases progressively with development. This increase in cell number is usually associated with the emergence of a "mature" zonation pattern. The typical internal structure of the shoot apex does not develop until a specific number of leaves form.

Gradual structural change in the growing point, however, does not adequately account for all aspects of juvenility. Sometimes, the transition from juvenile to adult leaf form is not graded but sudden. The juvenile leaves of species of the gymnosperm *Chamaecyparis,* for example, are needlelike and spreading; the adult leaves are scalelike and lie close to the stem. Among flowering plants, various species of *Eucalyptus* have juvenile leaves that are ovate and mature leaves that are sickle-shaped.

Phase change

Such sudden transitions from juvenile to adult form, referred to as phase change, seem to depend not on slow shifts in the apex but on some determinative event or correlated group of events. The two forms are relatively stable and tend to resist change; for example, cultured tissues taken from the juvenile (ivy-leaved) parts of ivy plants maintain a higher rate of cell division, and portions, or cuttings, taken from these parts tend to form roots more readily than those from the adult (simple-leaved) parts.

The establishment of these relatively stable but not wholly irreversible states is comparable with the determination of shoot and root poles during embryogenesis and, indeed, with the alternation of generations itself. The transmission of differentiated states through cell lineages presumably reflects the action of "switching" devices controlling the expression of different parts of the genetic complement. In this sense, phase change and related phenomena do not differ essentially from those of differentiation and organogenesis in general.

The transition in plants to the reproductive state is an example of a developmental event with some of the characteristics of phase change. Among seed plants, the reproductive structures are transformed shoots—strobili (including cones) of various kinds in the gymnosperms and flowers in angiosperms.

Flower development

From a developmental point of view, the flower can be regarded as a shoot axis of determinate growth, with the lateral members occupying the sites of leaves differentiating as floral organs—sepals, petals, stamens, and pistils. In the transition to flowering, the stem apex undergoes distinctive changes, the most conspicuous of which is in the shape of the apical region, which is related to the kind of structure to be formed, whether a single flower, as in the tulip, or a cluster of flowers (an inflorescence), as in the lilac. The region of cell division extends over the entire apex, and the ribonucleic acid content of terminal cells increases. When a single flower forms, lateral primordia emerge at higher and higher levels on the flanks of the apical dome, and the entire apex is absorbed in the process, after which apical growth ceases. When an inflorescence forms, early changes are generally comparable to that for the single flower with one major difference—axillary primordia emerge that either become floral meristems or develop as secondary inflorescence branches. These primordia appear closer to the apex than do those of axillary buds on a vegetative shoot. In grasses, the activation of axillary meristems is the most notable early indication of the passage into flowering.

The rate of maturation and the timing of the transition to the reproductive phase are sometimes governed by internal controls and thus are relatively insensitive to the environment, provided conditions are generally favourable for growth. Frequently, however, the developmental rate is affected profoundly by recurring cycles in the environment, particularly those of temperature and of day length. In effect, these cycles provide a timetable for the plant, thus adjusting flowering, fruiting, and seed dispersal to the season and increasing the chances for successful propagation.

The control of the developmental rate by temperature is especially evident in many herbaceous plants of temperate climates. These plants, as indicated earlier, often must experience cold, either as seeds or as young plants, before they can begin flower production; otherwise they undergo an excessively long period of leafy, or vegetative, growth.

Vernalization of the seed

After the cold experience, which can be given artificially, the plant is said to have been vernalized, or brought to the spring condition. Again the response is akin to a determination, because the condition attained is transmitted through subsequent cell divisions. Furthermore, there are indications that vernalization induces a persistent modification in the metabolism of apical cells and their derivatives. Ingenious theoretical schemes, offered to explain the apparent paradox that low temperature should actually accelerate a developmental process, are based mostly upon the proposition that a special vernalization hormone (vernalin) is involved. Although little direct evidence for the existence of vernalin exists, a class of hormones found in certain plant species, the gibberellins, does participate. The cold requirements of some species, such as the carrot, can be eliminated by the application of gibberellin, although the amounts needed are substantial.

Photoperiod as a trigger for flowering

The annual cycle of changing day length obviously provides the best of all "clocks" for the regulation of plant development. The effect of day length (or rather length of continuous darkness) on the transition to flowering is part of the general phenomenon of photoperiodism. Certain plants, called short-day plants, grow vegetatively when the nights are shorter than a critical minimum period (days long); exposure to longer nights (days short), however, accelerates development and brings on early flowering. Conversely, long-day plants develop very slowly toward flowering during daily cycles with longer than a minimum

of darkness (days short), and are accelerated by exposure to short nights (days long). Other plants either require days of intermediate length for flowering or respond to a sequence of different photoperiods.

The leaf, rather than the stem apex, is the light-receiving organ in the photoperiodic reaction, although it is at the apex that subsequent developmental changes occur. One commonly accepted view is that, as a consequence of the photoperiodic experience, a specific flower-inducing hormone (as yet not isolated but referred to as "florigen") is synthesized in the leaf and translocated to the apex. As in the case of vernalization, photoperiod undoubtedly affects the metabolism of the known plant hormones, and so influences many other developmental responses apart from flowering. The effect of the duration of illumination on the carbohydrate balance of the plant may also be important. Nutritional effects on flowering are well known in many species—certain fruit trees, for example.

Whether or not environmental factors influence the passage into a reproductive state of a plant, the transition must be viewed as part of the general development from juvenility to maturity: in this sense, flowering is not a radical alternative to vegetative growth but its culmination. Yet, entirely new organ types are produced at the flowering apex, presumably under the influence of genes inactive during vegetative growth.

SEASONAL ADAPTATIONS

Certain plants are perennial and survive from year to year by matching their growth to the progression of the seasons or by suspending growth altogether during unfavourable times, such as winter or a dry season.

Response to an approaching unfavourable season

In the temperate zone, some time before winter begins, growth ceases in the shoots of woody plants, resting buds are formed, and deciduous trees lose their leaves. The resting bud consists of a short axis, with the stem apex surrounded by modified unexpanded leaves, which protect the stem, especially from drying. The cells show marked frost resistance, similar to that of the embryo of the seed. Corresponding changes occur in herbaceous plants, in which the preparation for winter may involve the dying back of aerial parts altogether, leaving protected organs at or below the soil surface.

Growth sometimes ceases, even under favourable conditions, as a result of internal changes in the plant. This is true for some trees, which cease growth in midsummer. The passage into winter dormancy, however, is often controlled by the shortening of day length at the end of the growing season; in some plants decreasing night temperature also plays a part. Most temperate zone trees cease growth and form resting buds when the day length falls below a critical minimum.

Photoperiodic control seems to involve the formation of inhibitor compounds. In birches, for example, the leaf perceives the day length "signal" and transmits inhibitory materials to the apex, thus bringing growth to a stop and inducing the formation of a resting bud. The dormancy hormone, abscisic acid, may be concerned in this response and also in leaf abscission.

Budbreak

Budbreak in certain trees is controlled by photoperiod, growth resuming in the lengthening days of spring; light-perceptive organs are probably the young leaves inside the bud scales. Sometimes budbreak depends only on temperature increases that occur in spring, as in certain plants of Mediterranean climates.

The resumption of development in buds may result from a change in the balance of growth-inhibiting substances, such as abscisic acid, and growth promoters, notably the gibberellins. Buds can be caused to open prematurely by gibberellin treatment, which, as in the case of vernalization, can sometimes replace a cold experience; moreover, the gibberellin content in the buds of certain woody plants increases during chilling. Other hormones are probably also involved, however, for budbreak in plants such as the grapevine can be promoted by cytokinins, the plant cell-division factors.

An important general feature of adaptive periodicities is that the developmental changes anticipate the conditions for which they will ultimately provide the appropriate physiological or morphological adjustment. The ability of plants to utilize environmental indicators such as temperature and day-length changes is vital for the survival of plants. The production of such adaptive devices is made possible by the state of continuous embryogeny, already stressed as one of the most important characteristics of plant growth. (J.H.-H.)

ANIMAL DEVELOPMENT

General features

Development, in the context of this section, includes the processes that lead eventually to the formation of a new animal starting from cells derived from one or more parent individuals. Development thus occurs following the process by which a new generation of organisms is produced by the parent generation.

REPRODUCTION AND DEVELOPMENT

Metazoan reproduction

In multicellular animals (Metazoa), reproduction takes one of two essentially different forms: sexual and asexual. In asexual reproduction the new individual is derived from a blastema, a group of cells from the parent body, sometimes, as in *Hydra* and other coelenterates, in the form of a "bud" on the body surface. In sponges and bryozoans, the cell groups from which new individuals develop are formed internally and may be surrounded by protective shells; these bodies, which may serve as resistant forms capable of withstanding unfavourable environmental conditions, are released after the death of the parent. In certain animals the parent may split in half, as in some worms, in which an individual worm breaks into two fairly equal parts (except that the anterior half receives the mouth, "brain," and sense organs if they are present).

Obviously, in such a case it is impossible to say which of the two resulting individuals is the parent and which the offspring. Some brittle stars (starfish relatives) may reproduce by breaking across the middle of the body disk, with each of the halves subsequently growing its missing half and the corresponding arms.

A common feature of all forms of asexual reproduction is that the cells—always a substantial number of cells, never only one cell—taking part in the formation of the new individual are not essentially different from other body, or somatic, cells. The number of chromosomes (bodies carrying the hereditary material) in the cells participating in the formation of a blastema is the same as in the other somatic cells of the parent, constituting a normal, double, or diploid ($2n$), set.

In sexual reproduction, a new individual is produced not by somatic cells of the parent but by sex cells, or gametes, which differ essentially from somatic cells in having undergone meiosis, a process in which the number of chromosomes is reduced to one-half of the diploid ($2n$) number found in somatic cells; cells containing one set of chromosomes are said to be haploid (n). The resulting sex cells thus receive only half the number of chromosomes present in the somatic cell. Furthermore, the sex cells are generally capable of developing into a new individual only after two have united in a process called fertilization (see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The process of fertilization*).

Advantages of sexual and asexual reproduction

Each type of reproduction—asexual and sexual—has advantages for the species. Asexual reproduction is, at least in some cases, the faster process, leading most rapidly to the development of large numbers of individuals. Males and females are independently capable of producing offspring. The large size of the original mass of living matter and its high degree of organization—the new individual inherits parts of the body of the parent: a part of the alimentary canal, for instance—make subsequent development more

simple, and the attainment of a stage capable of self-support easier. New individuals produced by asexual reproduction have the same genetic constitution (genotype) as their parent and constitute what is called a clone. Though asexual reproduction is advantageous in that, if the parent animal is well adapted to its environment and the latter is stable, then all offspring will benefit, it is disadvantageous in that the fixed genotype not only makes any change in offspring impossible, should the environment change, but also prevents the acquisition of new characteristics, as part of an evolutionary process. Sexual reproduction, on the other hand, provides possibilities for variation among offspring and thus assists evolution by allowing new pairs of genes to combine in offspring. Since all body cells are derived from the fertilized egg cell, a mutation, or change, occurring in the sex cells of the parents immediately provides a new genotype in each cell of the offspring. In the course of evolution, sexual reproduction has been selected for, and established in, all main lines of organisms; asexual reproduction is found only in special cases and restricted groups of organisms.

PREPARATORY EVENTS

In the case of multicellular animals we find there are two kinds of sex cells: the female sex cell (ovum, or egg), derived from an oocyte (immature egg), and the male sex cell (spermatozoon or sperm), derived from a spermatocyte. Eggs are produced in ovaries; sperm, in testes. Both the egg and the sperm contribute to the development of the new individual; each providing one set of genes, thereby restoring the diploid number of chromosomes in the fertilized egg. The sperm possesses a whiplike tail (flagellum) that enables it to swim to the egg to fertilize it. In most cases the egg, a stationary, spherical cell, provides the potential offspring with a store of food materials, or yolk, for its early development (Figure 12). The term yolk does not

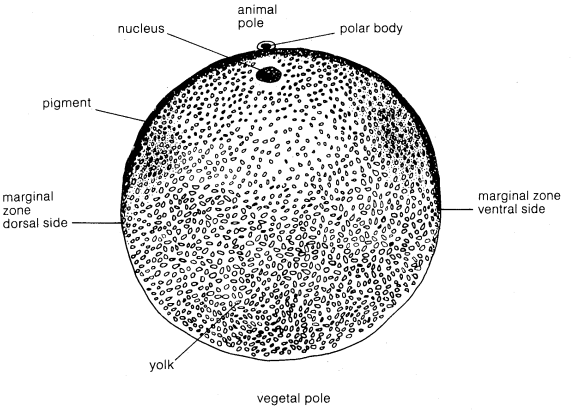


Figure 12: Amphibian egg, vertical section, showing distribution of yolk and pigment in interior.

refer to any particular substance but in fact includes proteins, phosphoproteins, lipids, cholesterol, and fats, all of which substances occur in various proportions in the eggs of different animals. In addition to yolk, eggs accumulate other components and acquire the structure necessary for the development of the new individual. In particular the egg acquires polarity—that is, the two ends, or poles, of the egg become distinctive from each other. At one pole, known as the animal pole, the cytoplasm appears to be more active and contains the nucleus (meiotic divisions occur in this region); at the other, called the vegetal pole, the cytoplasm is less active and contains most of the yolk. The general organization of the future animal is closely related to the polarity of the egg.

When the amount of food reserve is comparatively small, as it is in many marine invertebrates and mammals (in the latter the embryo is nourished by materials in the mother's blood), the egg may be barely visible to the unaided eye. The egg of the sea urchin is about 75 microns (0.003 inch) in diameter; that of a human being is slightly more than 0.1 millimetre. Eggs are classified according to the amount of yolk present. An egg with a small quantity of evenly distributed yolk is called an oligolecithal egg. One with more yolk that is unevenly distributed (*i.e.,* concentrated towards the vegetal pole) is telolecithal; and one with still greater amounts of yolk in granules or in a compact mass is megalecithal.

Protective membranes

The egg is surrounded by protective membranes, which may be soft and jellylike or hard and calcified, like shells. Egg membranes are produced while the egg is either in the ovary or being carried away from the ovary in a tube called an oviduct. The eggs of many animals have both kinds of membranes. In insects, a hard shell (chorion) forms around the eggs in the ovaries. In frogs, a very thin vitelline membrane forms around the eggs in the ovary; subsequently a layer of jelly is deposited around the eggs while they pass through the oviducts. In birds, a very thin vitelline membrane is produced around the egg in the ovary; then several layers of secondary membranes are formed in the oviduct before the egg is laid. The outermost of these secondary membranes is the calcareous shell. In mammals the egg is surrounded by the so-called pellucid zone, which is equivalent to the vitelline membrane of other animals; follicle cells form an area called the corona radiata around this zone.

After fertilization the egg, now called a zygote, is endowed with genes from two parents and has begun actual development. (Activation of the egg may be brought about by an agent other than sperm in certain animals, but such cases of parthenogenesis are exceptional. (See REPRODUCTION AND REPRODUCTIVE SYSTEMS.)

After fertilization, the zygote undergoes a series of transformations that bring it closer to the essential organization of the parents. These transformations, initiated at a physiological, perhaps even at a molecular, level, eventually result in the appearance of certain structures. The whole process is called morphogenesis (Greek *morphē,* "shape" or "form"; *genesis,* "origin" or "production"). The process of development is more easily understood if, at every step, the changes necessary to bring the system nearer the goal are considered. Depending on the achievements necessary at any step, development can be subdivided into a number of discrete phases, the first of which, cleavage, immediately follows fertilization.

Early development

EMBRYO FORMATION

Cleavage. Since the goal of development is the production of a multicellular organism, many cells must be produced from the single-celled zygote. This task is accomplished by cleavage, a series of consecutive cell divisions. Cells produced during cleavage are called blastomeres. The divisions are mitotic—*i.e.,* each chromosome in the nucleus splits into two daughter chromosomes, so that the two daughter blastomeres retain the diploid number of chromosomes. During cleavage, almost no growth occurs between consecutive divisions, and the total volume of living matter does not change substantially; as a consequence, the size of the cells is reduced by almost half at each division. At the beginning of cleavage, cell divisions tend to occur at the same time in all blastomeres, and the number of cells is doubled at each division. As cleavage progresses, the cells no longer divide at the same time.

Cleavage in most animals follows an orderly pattern, with the first division being in the plane of the main axis of the egg (Figure 13). This cleavage plane is arbitrarily called vertical, on the assumption that the main axis of the egg is vertical. The second cleavage plane is again vertical but at right angles to the first, giving rise to four equal cells arranged around the main axis of the egg (Figure 13). The third cleavage plane is at right angles to both the first and second cleavage planes and is horizontal, or equatorial. Subsequent divisions may alternate between vertical and horizontal cleavage planes, but later cleavage divisions become randomly oriented. This pattern is typical of many animal groups; however, more complicated patterns of cleavage are found in such animals as annelids, mollusks, and nematodes.

As the amount of yolk in the egg increases, it influences cleavage by hindering the cytoplasmic movements

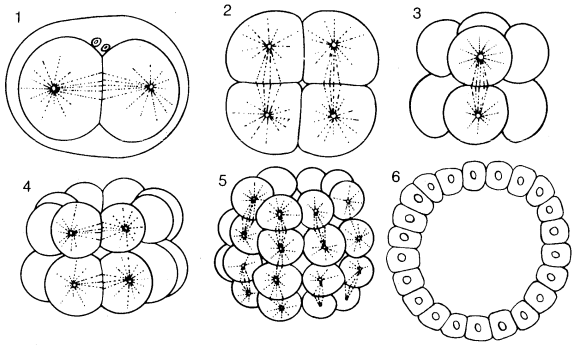


Figure 13: Cleavage of an oligolecithal egg (sea cucumber) in progressive stages 1 through 6 (see text).

Influence of yolk on cleavage

involved in mitosis. If there is only little yolk (oligolecithal eggs), the yolk granules follow the movements of the cytoplasm and are distributed in the resulting blastomeres. But if the amount of yolk is larger (megalecithal eggs), cleavages occur nearer the animal pole, where there is less yolk; as a result, the blastomeres nearer the animal pole are smaller than those nearer the vegetal pole. The presence of yolk masses may retard the onset of cleavage in a part of the egg or even suppress it altogether; in this case cleavage is partial, or meroblastic. Only a part of the egg material then is subdivided into cells, the rest remaining as a mass that serves as nourishment for the developing embryo.

Cleavage is complete, or holoblastic, in many invertebrates including coelenterates, annelids, echinoderms, tunicates, and cephalochordates. The blastomeres may be either about equal or only slightly different in size. Cleavage in amphibians is holoblastic, but the size of the blastomeres is very uneven. Blastomeres are smallest at the animal pole and largest (and yolky) at the vegetal pole. Somewhat similar conditions prevail in many mollusks. In most fishes, birds, reptiles, and egg-laying mammals (monotremes), cleavage is discoidal—*i.e.*, restricted to a disk of cytoplasm at the animal pole of the egg, most of the yolky egg material remaining uncleaved (Figure 14). Cleavage in insects and many other arthropods is

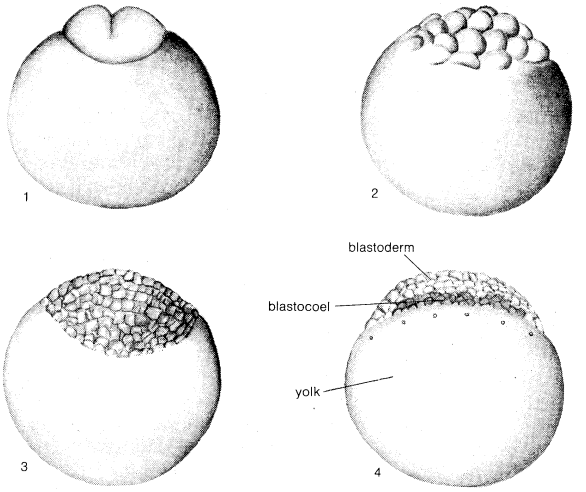


Figure 14: Discoidal cleavage of an egg of a bony fish.

superficial—*i.e.*, the entire surface layer of egg cytoplasm subdivides into cells, and the egg contains a central mass of uncleaved yolk. The conditions of cleavage in placental mammals, including man, are peculiar.

During cleavage, development involves only an increase of cell numbers; the shape of the embryo does not change, and chemical transformations within the embryo are restricted to those necessary for cell division. Chemical and structural transformations are concerned with accumulating chromosomal material in the nuclei of the blastomeres.

Chromosome doubling

Before each division the chromosomes carrying the genes double in number; this means that the chromosomal material, deoxyribonucleic acid (DNA), has to be synthesized. This synthesis proceeds possibly at the expense of cytoplasmic ribonucleic acid (RNA) but certainly also from simpler organic compounds. A certain amount of protein synthesis is also necessary for cleavage to proceed: if developing eggs are treated with puromycin, a substance which is known to suppress protein synthesis, cleavage stops immediately. The proteins concerned have not yet been identified. No proteins are synthesized, however, that would foreshadow the future differentiation of parts of the embryo. It is believed that the genes in the chromosomes remain largely inactive during cleavage. The rhythm (speed) of cleavage is wholly dependent on the cytoplasm of the egg.

Although the shape and volume of the embryo do not change during cleavage, one important change in gross organization does take place. As the blastomeres are produced, they move outward, leaving a centrally located fluid-filled cavity. In cases of holoblastic cleavage, the blastomeres become arranged in a layer from one to several cells thick surrounding the cavity. The embryo at this stage may be likened to a hollow ball and is known as a blastula (Figure 15:1). The outer layer of cells is called the blastoderm, and the fluid-filled cavity the blastocoel. In discoidal cleavage the cells, which do not surround the whole embryo, lie only on the animal pole; nevertheless, a blastocoel may be formed by a crevice appearing between the blastomeres and the mass of yolk (Figure 14). The blastomeres then may be arranged as a saucer-shaped blastodisk covering the blastocoel.

The formation of the blastula signifies the end of the period of cleavage. The next stage of development is concerned not with an increase in cell number, though cell divisions continue at a slower pace, but with rearrangement of the available cell masses to conform with the gross features of the future animal.

Gastrulation. The embryo in the blastula stage must go through profound transformations before it can approach adult organization. An adult multicellular animal typically possesses a concentric arrangement of tissues of the body; this feature is common to all animal groups above the level of the sponges. Adult tissues are derived from three embryonic cell layers called germinal layers: the outer layer is the ectoderm, the middle layer is the mesoderm, and the innermost layer is the endoderm (entoderm). The ectoderm gives rise to the skin covering, to the nervous system, and to the sense organs. The mesoderm produces the muscles, excretory organs, circulatory organs, sex organs (gonads), and internal skeleton. The endoderm lines the alimentary canal and gives rise to the organs associated with digestion and, in chordates, with breathing.

Germinal layers

The blastula, which consists of only one cell layer, undergoes a dramatic reshuffling of blastomeres preparatory to the development of the various organ systems of the animal's body. This is achieved by the process of gastrulation, which is essentially a shifting or moving of the cell material of the embryo so that the three germinal layers are aligned in their correct positions.

The rearrangement of the blastula to form the germinal layers is seen clearly in certain marine animals with oligolecithal eggs. The hollow blastula consists of a simple epithelial layer (the blastoderm), the transformation of which can be likened to the pushing in of one side of a rubber ball (Figure 15). As a result of such inpushing (or invagination), the spherical embryo is converted into a double-walled cup, the opening of which represents the position of the former vegetal pole. The involuted part of the blastoderm, lining the inside of the double-walled cup, gives rise to the endoderm and mesoderm, and the blastomeres remaining on the exterior become the ectoderm. As a consequence of the infolding at the vegetal pole, the blastocoel is reduced or obliterated, and a new cavity is created, the primitive gut cavity, or archenteron, which eventually gives rise to the hollow core (lumen) of the alimentary canal. At this stage the embryo has a primitive gut with an opening to the exterior and is known as a gastrula. The opening of the gastrula is the blastopore, or primitive mouth; both terms are somewhat misleading. It would seem that the term blastopore should be applied more appropriately to an opening in a blastula, in which, of course, no opening exists. As to the term primitive

The gastrula stage

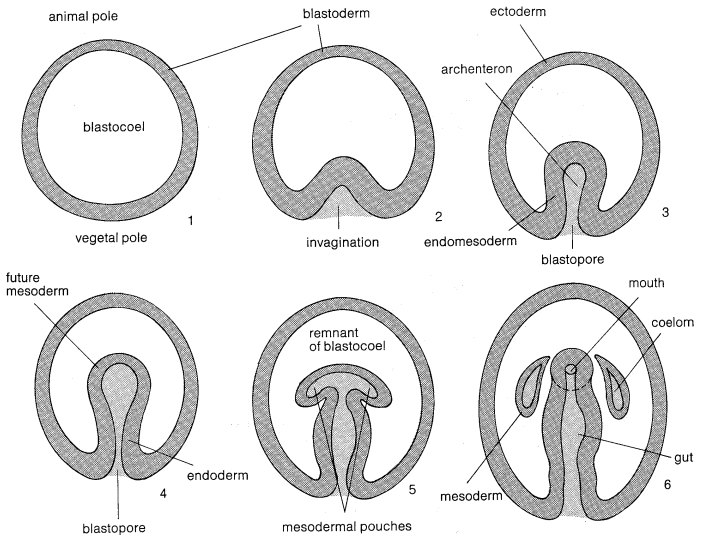


Figure 15: Gastrulation and mesoderm formation in the embryo of a starfish.

mouth, it must be pointed out that the blastopore does not always give rise to the adult mouth. In certain animal groups it becomes the anus, and a mouth forms as a completely new opening.

In some coelenterates, cells at the vegetal pole do not form an invaginating pocket, but individual cells slide inward, losing connection with other cells of the blastoderm. Eventually these cells fill the blastocoel and form a compact mass of endoderm. The cavity (archenteron) within this mass and the opening (blastopore) to the exterior are then produced secondarily by the separation of these cells.

Amphioxus, echinoderms, and amphibians. Gastrulation does not always proceed exactly as described above. In the course of evolution, certain animal groups have modified this critical stage of embryonic development, and these modifications have undoubtedly contributed to the successful continuation of species. In the primitive fish-like chordate amphioxus, for example, the invaginating blastoderm eventually comes into close contact with the inner surface of the ectoderm, thus practically squeezing the blastocoel out of existence or at least reducing it to a narrow crevice between the ectoderm and the endomesoderm. In echinoderms, on the other hand, a smaller portion of the blastoderm invaginates, and the blastocoel remains as a spacious internal cavity between the ectoderm and the endomesoderm. It persists as the primary body cavity and is the only body cavity (apart from the cavity of the alimentary canal) in such invertebrates as nematodes and rotifers.

In the double-walled-cup stage, the two internal germinal layers—endoderm and mesoderm—may not yet be distinct. Their separation may occur later, in the second phase of gastrulation, by one of two methods. One is the development of outpocketings from the wall of the archenteron. In starfishes and other echinoderms, the deep part of the endomesodermal invagination forms two thin-walled sacs, one on each side of the gastrula (Figure 15). These are the rudiments of the mesoderm; the remaining part of the archenteron becomes the endoderm and produces the lining of the gut. The cavities within the mesodermal sacs expand to become the coelom, the secondary body cavity of the animal. A somewhat similar process of mesoderm and coelom development occurs in amphioxus among the chordates, except that a series of mesodermal sacs forms on either side of the embryo, foreshadowing the segmented (metameric) structure common to chordates. Only the most anterior pairs of the mesodermal sacs actually contain a cavity at the time of their formation; the more posterior ones are solid masses of cells separating from the archenteric wall and from one another and developing coelomic cavities later.

Gastrulation in starfishes

A second method of mesoderm formation is by the splitting off of mesodermal cells from the original common mass of endomesoderm. This may take the form of single cells detaching themselves from the archenteron or of whole sheets of cells splitting off from the endoderm. An example of the latter type is seen in the gastrulation of amphibians. The development of specific regions of the early amphibian embryo—by the use of natural pigmentation or artificially introduced dyes—can be followed and their location in the adult recorded in diagrams called fate maps. The fate map of a frog blastula just prior to gastrulation demonstrates that the materials for the various organs of the embryo are not yet in the position corresponding to that in which the organs will lie in a fully developed animal. The endodermal material for the foregut, for example, lies not far from the vegetal pole; the ectodermal component of the mouth region (stomodeum) is situated close to the animal pole. Extensive rearrangement of the embryo is necessary to bring all the parts into their correct relationships.

Because of the large amount of yolk and resulting uneven cleavage, gastrulation in amphibians cannot proceed by a simple infolding of the vegetal hemisphere. A certain amount of invagination does take place, assisted by an active spreading of the animal hemisphere of the embryo; as a result, the ectoderm covers the endodermal and mesodermal areas. The spreading is sometimes described as an "overgrowth"—an inappropriate term, since no growth or increase of mass is involved. The future ectoderm simply thins out, expands, and covers a greater surface of the embryo in a movement known as epiboly.

Gastrulation in amphibians, in lungfishes, and in the cyclostomes (hagfishes and lampreys) begins with the formation of a pit on what will become the back (dorsal) side of the embryo (Figure 16). The pit represents the active

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

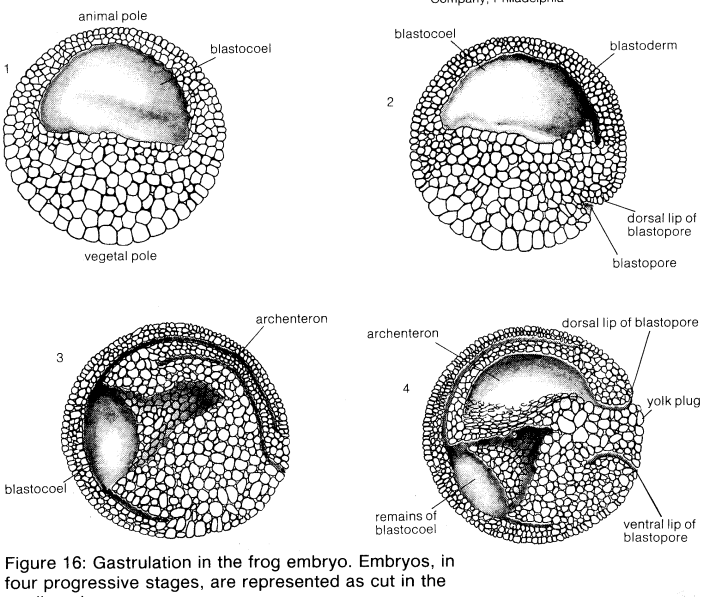


Figure 16: Gastrulation in the frog embryo. Embryos, in four progressive stages, are represented as cut in the median plane.

shifting inward of the cells of the blastoderm. As these cells undergo a change in shape, there occurs also a contraction at the external surface, with adjacent cells being drawn toward the centre of the contraction even before an actual depression is formed. The cells most concerned in this process will become part of the future foregut. Further movement of the cells inward results in the formation of a distinct pit, which rapidly develops into a pocket-like archenteron with its opening, the blastopore. Once the archenteron is formed, more and more of the exterior cells roll over the edge of the blastopore and disappear into the interior. In the course of gastrulation the shape of the blastopore changes from a simple pit to a transverse slit and finally into a groove encircling the yolky material at the vegetal pole. As a result of epiboly of the animal hemisphere, the upper edge of the groove is gradually pushed down until the yolky cells of the vegetal pole are covered completely. The edges of the blastopore then converge toward the vegetal pole, the slit between them

Formation of archenteron in amphibians

being eventually reduced to a narrow canal, which lies at the posterior end of the embryo and, in some species, becomes the anal opening. (In other cases the canal closes, and a new anal opening breaks through nearby, slightly more ventrally.)

The cavity of the archenteron increases as more material from the outside is transferred inward, and the blastocoel becomes almost completely obliterated. Both mesoderm and endoderm are shifted into the interior, and only the ectoderm remains on the embryo surface. The mesoderm splits from the endoderm: the endoderm lines the archenteric cavity (and eventually becomes the lining of the alimentary canal), as the mesoderm surrounds the endoderm to form the chordamesodermal mantle. By the time the blastopore closes, the three germ layers are in their correct spatial relationship to each other.

Reptiles, birds, and mammals. Although amphibian gastrulation is considerably modified in comparison with that in animals with oligolecithal eggs (*e.g.,* amphioxus and starfishes), an archenteron forms by a process of invagination. Such is not the case, however, in the higher vertebrates that possess eggs with enormous amounts of yolk, as do the reptiles, birds, and egg-laying mammals. Cleavage in these animals is partial (meroblastic), and, at its conclusion, the embryo consists of a disk-shaped group of cells lying on top of a mass of yolk. This cell group often splits into an upper layer, the epiblast, and a lower layer, the hypoblast. These layers do not represent ectoderm and endoderm, respectively, since almost all the cells that form the embryo are contained in the epiblast. Future mesodermal and endodermal cells sink down into the interior, leaving only the ectodermal material at the surface. In reptiles, egg-laying mammals, and some birds, a pocket-like depression occurs in the epiblast but encompasses only chordamesoderm or even only the notochord. Individual cells of the remainder of the mesoderm and endoderm migrate into the interior and there arrange themselves into a sheet of chordamesoderm and of endoderm, the latter of which mingles with cells of the hypoblast if such a layer is present. The migration of the cells destined to form mesoderm and endoderm does not take place over the whole surface of the disk-shaped embryo but is restricted to a specific area along the midline. This area is more or less oval in reptiles and lower mammals; distinctly elongated in higher mammals and birds, it is called the primitive streak (Figure 17), a thickened and slightly depressed part of the epiblast that is thickest at the anterior end, called the Hensen's node.

In animals having discoidal cleavage, the three germinal layers at the end of gastrulation are stacked flat; ectoderm on top, mesoderm in the middle, and endoderm at the bottom. The embryo is produced from the flattened layers by a process of folding to form a system of concentric tubes (Figure 18). The edges of the germ layers, which are not involved in the folding process, remain attached to the yolk and become the extra-embryonic parts; they are not directly involved in supplying cells for the embryo but break down yolk and transport it to the developing embryo.

Higher mammals—apart from the egg-laying mammals—do not have yolk in their eggs but, having passed through an evolutionary stage of animals with yolky eggs, retain,

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

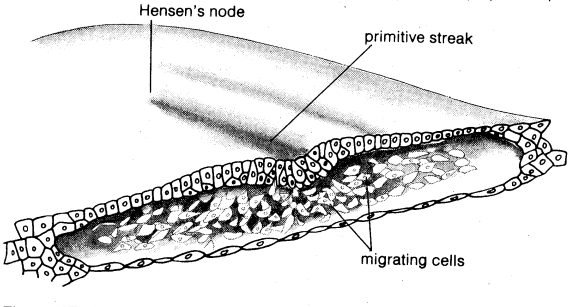


Figure 17: Anterior half of blastoderm of a bird embryo cut transversely to show migration of mesodermal and endodermal cells from the primitive streak.

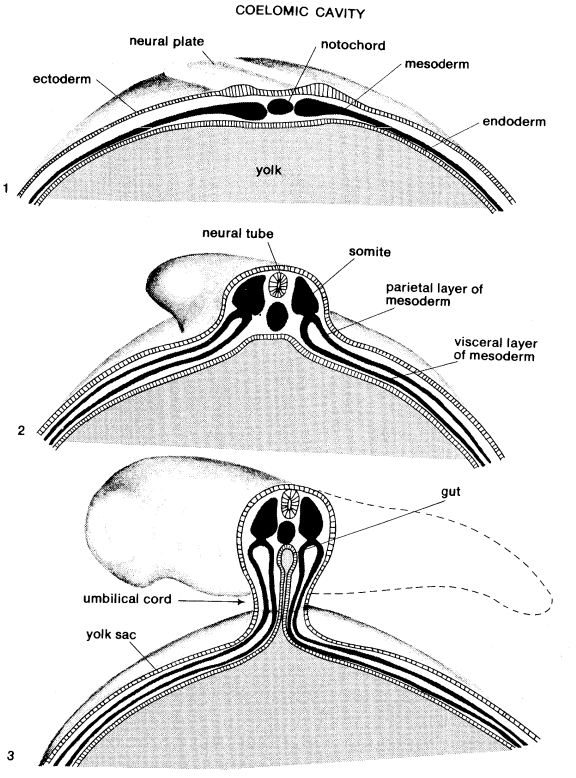


Figure 18: Vertebrate body formation from three germinal layers showing separation from yolk sac after discoidal cleavage.

Gastrulation in higher mammals

particularly in gastrulation, features common to reptiles (and birds, which also had reptilian ancestors). As a result, at the end of cleavage the formative cells of the embryo—the cells that will actually build the body of the animal—are arranged in the form of a disk over a cavity that takes the place of the yolk of the reptilian ancestors of mammals. Within the disk of cells a primitive streak develops, and the three germinal layers are formed much as in many reptiles and birds.

Gastrulation and the formation of the three germinal layers is the beginning of the subdivision of the mass of embryonic cells produced by cleavage. The cells then begin to change and diversify under the direction of the genes. The genes brought in by the sperm exert control for the first time; during cleavage all processes seem to be under control of the maternal genes. In cases of hybridization, in which individuals from different species produce offspring, the influence of the sperm is first apparent at gastrulation: paternal characteristics may appear at this stage; or the embryo may stop developing and die if the paternal genes are incompatible with the egg (as is the case in hybridization between species distantly related).

The diversification of cells in the embryo progresses rapidly during and after gastrulation. The visible effect is that the germinal layers become further subdivided into aggregations of cells that assume the rudimentary form of various organs and organ systems of the embryo. Thus the period of gastrulation is followed by the period of organ formation, or organogenesis.

EMBRYONIC ADAPTATIONS

Throughout its development the embryo requires a steady supply of nourishment and oxygen and a means for disposal of wastes. These needs are met in various ways, depending in particular on (1) whether the eggs develop externally (oviparity), are retained in the maternal body until ready to hatch (ovoviviparity), or are carried in the maternal body to a later stage (viviparity); and (2) the length of embryonic development.

Adaptations in animals other than mammals. Eggs of many marine invertebrates are discharged directly into water, and the period of development before the larva emerges is relatively brief. Oxygen diffuses easily into the

small eggs, and nourishment is provided by a moderate amount of yolk. During cleavage the yolk is distributed to all the blastomeres. Much of the nourishment in the egg is stored as animal starch, or glycogen, which is almost completely used by the time the larva emerges from the egg. A small amount of water and inorganic salts are taken in by the embryo from surrounding seawater. Eggs developing in freshwater carry their own supply of necessary amounts of certain salts that are not present in sufficient quantities in the environment. Products of metabolism—especially carbon dioxide and nitrogenous wastes in the form of ammonia—diffuse out from small embryos developing in water.

Mechanisms to overcome drying

The eggs of terrestrial animals must overcome the hazard of drying. In certain species this danger is avoided because the animal returns to water to breed, such as frogs and salamanders. Some groups of insects (*e.g.*, dragonflies, mayflies, and mosquitoes) also lay eggs in water, and the larvae are aquatic. Eggs of other animals (*e.g.*, snails, earthworms) are laid in moist earth and thus are protected from drying up. In terms of evolution, however, a decisive solution to the problem of development on land was arrived at by most insects and by reptiles and birds, which developed eggs with a shell impermeable to water or, at least, resistant to rapid evaporation. The shells of bird and insect eggs, while restricting evaporation of water, allow oxygen to diffuse into the egg and carbon dioxide to diffuse out. Apart from gas exchange, the eggs constitute closed systems, which give nothing to the outside and require nothing from it. Such eggs are called cleidoic. Because the products of nitrogen metabolism in cleidoic eggs cannot pass through the eggshell, animals (birds and insects) have had to evolve a method of storing wastes in the form of uric acid, which, since it is insoluble, is nontoxic to the embryo.

After a short period of development in the egg, the emerging young animal has to fend for itself, unless there is some form of parental care. Exposure to the external environment at a tender age results frequently in loss of life, a hazard met by many animals through an increase in the supply of nourishment within the egg, thus allowing the young to attain a greater size and development. This tendency to produce large yolky eggs has been achieved independently in different evolutionary lines: in octopuses and squids among the mollusks, in sharks among the fishes, and in reptiles and birds among the terrestrial vertebrates.

As has been indicated, cleavage is incomplete in eggs with large amounts of yolk. Although some yolk platelets may be enclosed in the formative cells of the embryo, the bulk of the yolk remains an uncleaved mass, overgrown and surrounded by the cellular part of the embryo. In such cases a membranous bag, or yolk sac, is formed (Figure 19) and remains connected to the embryo by a narrow stalk (the evolutionary precursor of the umbilical cord of mammals). The cellular layers surrounding the yolk sac and forming its walls may consist of all three germinal layers (in reptiles and birds), so that the yolk virtually comes to lie inside an extension of the gut of the embryo; or (in bony fishes) the yolk sac may be enclosed in layers of ectoderm and mesoderm. In either case a network of blood vessels develops in the walls of the yolk sac and transports the yolk products to the embryo. As the yolk is broken down and utilized, the yolk sac shrinks and is eventually drawn into the body of the embryo. In addition to the yolk sac, extra-embryonic parts are also encountered in the form of embryonic membranes, which are found in higher vertebrates and in insects. Vertebrates have three embryonic membranes: the amnion, the chorion, and the allantois.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

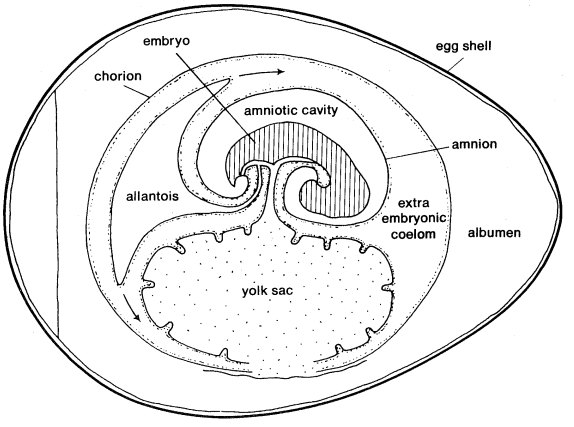


Figure 19: Position of a bird embryo in the egg and the relation of the allantois to the amniotic cavity and the yolk sac.

In reptiles, birds, and mammals, folds develop on the surface of the yolk sac just outside and around the body of the embryo proper. These folds, consisting of extra-embryonic ectoderm and extra-embryonic mesoderm, rise up and fuse dorsally, enclosing the embryo in a double-lined, fluid-filled chamber known as the amniotic cavity. The inner lining of the fold becomes the amnion, and the outer becomes the chorion, which ultimately surrounds the entire embryo. The amniotic fluid protects the embryo from drying, prevents the adhesion of the embryo to the inner surface of the shell, and provides the embryo with efficient shock absorption against possible damaging jolts. (The aminion and chorion develop in the same way in insect embryos.) The third membrane, or allantois, is originally nothing more than the urinary bladder of the embryo. It is a saclike growth of the floor of the gut, into which nitrogenous wastes of the embryo are voided. It enlarges greatly during the course of development, eventually expanding between the amnion and chorion and also between the chorion and the yolk sac, to become the third embryonic membrane. In addition to providing storage space for the nitrogenous wastes of the embryo, the allantois takes up oxygen, which penetrates into the egg from the exterior, and delivers it, by way of a network of blood vessels, to the embryo.

The allantois

Adaptations in mammals. At some early stage during the evolution of viviparous mammals, eggs came to be retained in the oviducts of the mother. The embryo then was provided with nourishment from fluids in the oviduct; the yolk, which became redundant, gradually ceased to be provided, and the eggs became oligolecithal. The eggshell, present in reptiles, was no longer needed and eventually disappeared, as did the white of the egg. The chorion, however, remained as the most external coat of the developing embryo through which nourishment reaches the embryo. It acquired the ability to adhere closely to the walls of the uterus (which was what that part of the oviduct holding the embryo had become) and became the so-called trophoblast. The blood-vessel network of the underlying allantois conveys nutrients that diffuse through the trophoblast to the body of the embryo proper. These modifications gave rise to a new organ, the placenta, formed from tissues of both the mother and the embryo: the uterine wall with its blood vessels provided by the mother; the trophoblast and allantois—and in some mammals also the yolk sac—with their blood vessels provided by the embryo.

The overall development of placental mammals as a result of these changes is profoundly different from that of their ancestors, the reptiles, and proceeds in the following way: the tiny yolkless egg is fertilized in the upper portion of the oviduct by sperm received from the male in the process of coupling (coitus); cleavage starts as the egg is propelled slowly down the oviduct by action of cilia in the oviduct lining. At the end of cleavage a solid ball of cells called a morula is produced. The surface cells of the morula become the trophoblast and the inner cell mass gives rise to the embryo (the formative cells) and also its yolk sac, amnion, and allantois. A cavity appears within the morula, converting it into a hollow embryo, called the blastocyst. This cavity resembles the blastocoel but, in fact, is analogous to the yolk sac of meroblastic eggs, except that there is no yolk and the cavity is filled with fluid. At the blastocyst stage, the embryo enters the uterus and attaches itself to the uterine wall. This attachment, or implantation, a crucial step in the development of a mammal, is attained through the action of the trophoblast,

Implantation

which forms extensions, known as villi, that penetrate the uterine wall (Figure 20). In higher placental mammals, the lining of the uterine wall and, in varying degrees, the underlying tissues as well are partially destroyed, resulting in a closer relationship between the blood supplies of the mother and the embryo. Indeed, in man and in some rodents, the blastocyst sinks completely into the uterine wall and becomes surrounded by uterine tissue.

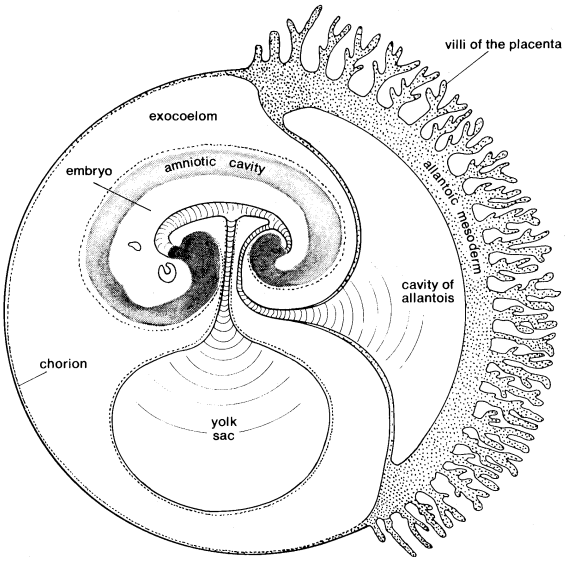


Figure 20: Embryo of placental mammal surrounded by embryonic membranes.

While implantation takes place, the formative cells arrange themselves in the form of a disk under the trophoblast. In the disk, the germinal layers develop much as in birds, with the formation of a primitive streak and migration of the chordamesoderm into a deeper layer. A layer of endoderm is formed adjoining the cavity of the blastocyst, and an amniotic cavity develops, enclosing the embryo; in lower placental mammals, the allantois also develops. The embryo proper, lying in the amniotic cavity, is connected to the extra-embryonic parts by the umbilical cord. The umbilical cord lengthens greatly during later development. In higher mammals, the cavity of the allantois is reduced, but the allantoic blood vessels become well developed and extend through the umbilical cord, connecting the embryo to the placenta. The blood that circulates in the placenta brings oxygen and nutrients from the maternal blood to the embryo and carries away carbon dioxide and other waste products from the embryo to the maternal blood for disposal by the maternal body.

Although tissues of maternal and embryonic origin are closely apposed in the placenta, there is little actual mingling of the tissues. Despite an occasional penetration of an embryo cell into the mother and vice versa, there is a placental barrier between the two tissues. The blood circulation of the mother is at all times completely separated from that of the embryo and its extra-embryonic parts. The placental barrier, however, does allow molecules of various substances to pass through; such differential permeability is indeed necessary if the embryo is to obtain nourishment. The permeability of the placental barrier differs in different animals; thus antibodies, which are protein molecules, may penetrate the placental barrier in man but not in cattle.

The maintenance of the fetus—as the more advanced embryo of a mammal is called—in the uterus is under hormonal control. In the initial stages of pregnancy, the continued existence of the embryo in the uterus depends on the hormone progesterone, which is secreted by the corpora lutea, "yellow bodies," that develop in the ovary after an egg has been released.

At birth the fetal parts of the placenta separate from the maternal parts. Contraction of the uterine wall first releases the fetus from the uterus; the fetal parts of the placenta (the afterbirth) follow. In certain cases of intimate connection between fetal and maternal tissues, the maternal tissues are torn, and birth is accompanied by profuse bleeding.

Organ formation

PRIMARY ORGAN RUDIMENTS

Immediately after gastrulation—and sometimes even while gastrulation is underway—the germinal layers begin subdividing into regions that will give rise to various parts of the body. Subdivision proceeds in stages: initially a mass of cells is set aside for an organ system (for the alimentary canal, for instance) and subsequently further subdivided into the rudiments of various parts of the organ system, such as the liver, stomach, and intestines. The initially formed larger units are referred to as primary organ rudiments; those they later give rise to, as secondary organ rudiments.

The type of organ rudiment produced depends on the organization of the body in any particular group in the animal kingdom. In the vertebrates the earliest subdivision within a germinal layer is the segregation within the chordamesodermal mantle of the rudiment of the notochord from the rest of the mesoderm (Figure 18). During gastrulation the material of the notochord comes to lie middorsally in the roof of the archenteron. It separates by longitudinal crevices from the chordamesodermal mantle lying to the left and right. The material of the notochord then rounds off and becomes a rod-shaped strand of cells immediately under the dorsal ectoderm, stretching from the blastopore toward the anterior end of the embryo, to the midbrain level. In front of the tip of the notochord, there remains a thin sheet of prechordal mesoderm.

Development of notochord

The mesodermal layer adjoining the notochord becomes thickened and, by transverse crevices, subdivided into sections called somites (Figure 21). The somites, which later give rise to the segmented body muscles and the vertebral column, are the basis of the segmented organization typical of vertebrates (seen especially in the lower fishlike forms but also in the embryos of higher vertebrates). The lateral and ventral mesoderm, which remains unsegmented, is called the lateral plate. The somites remain connected to the lateral plate by stalks of somites that play a particular role in the development of the excretory (nephric) system in vertebrates; for this reason they are called nephrotomes. Rather early the mesodermal mantle splits into two layers, the outer parietal (somatic) layer and the inner visceral (splanchnic) layer, separated by a narrow cavity (Figure 22) that will expand later to form the coelomic, or secondary, body cavity. The coelomic cavity extends initially through the nephrotomes into the somites; in the somites it is eventually obliterated. Endoderm completely surrounds the lumen of the archenteron (when present) and produces the cavity of the alimentary canal. If no archenteric cavity is formed during gastrulation, the cavity of the alimentary canal is formed by the separation of cells in the middle of the mass of endoderm (as in bony fishes) or by folding of the sheet of endoderm (Figure 18). The endodermal gut sooner or later acquires an extended anterior part called the foregut and a narrower and more elongated posterior part, the hindgut. Characteristic of chordates is the development of the nervous system from a part of ectoderm lying originally on the dorsal side of the embryo, above the notochord and the somites. This part of the ectodermal layer thickens and becomes the neural plate (Figure 18:1), whose edges rise as neural folds that converge toward the midline, fuse together, and form the neural tube. In vertebrates the neural tube lies immediately above the notochord and extends beyond its anterior tip. The neural tube is the rudiment of the brain and spinal cord; its lumen gives rise to the cavities, or ventricles, of the brain and to the central canal of the spinal cord. The remainder of the ectoderm closes over the neural tube and becomes, in the main, the covering layer (epithelium) of the animal's skin (epidermis). As the neural tube detaches itself from the overlying ectoderm, groups of cells pinch off and form the neural crest, which plays an important role in the development of, among other things, the segmental nerves of the brain and spinal cord.

Formation of the neural crest

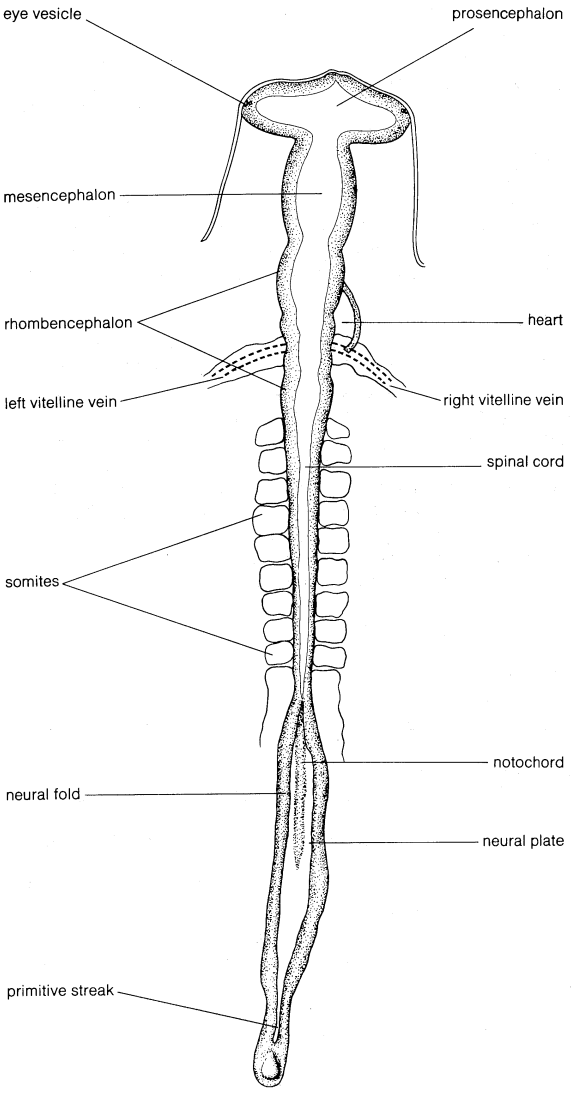


Figure 21: Early chick embryo, showing development of the brain vesicles and of some of the mesodermal structures.
From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

In developing the primary organ rudiments mentioned above, the embryo acquires a definite organization clearly recognizable as that of a chordate animal. Similar processes, which occur in the development of other animals, establish the basic organization of an annelid, a mollusk, or an arthropod.

The organization of the embryo as a whole appears to be determined to a large extent during gastrulation, by which process different regions of the blastoderm are displaced and brought into new spatial relationships to each other. Groups of cells that were distant from each other in the blastula come into close contact, which increases possibilities for interaction between materials of different origin. In the development of vertebrates in particular, the sliding of cells (presumptive mesoderm) into the interior and their placement on the dorsal side of the archenteron (in the archenteric "roof"), in immediate contact with the overlying ectoderm, is of major importance in development and subsequent differentiation. Experiments have shown that, at the start of gastrulation, ectoderm is incapable of progressive development of any kind; that only after invagination, with chordamesoderm lying directly underneath it, does ectoderm acquire the ability for progressive development. The dorsal mesoderm, which later differentiates into notochord, prechordal mesoderm, and somites, causes the overlying ectoderm to differentiate as neural plate. Lateral mesoderm causes overlying ectoderm to differentiate as skin. The influence exercised by parts of the embryo, which causes groups of cells to proceed along a particular path of development, is called embryonic induction. Though induction requires that the interacting parts come into close proximity, actual contact is not necessary. The inducing influence—whatever it might be—is a diffusible substance emitted by the activating cells (the inductor). The inducing substance of the mesoderm is a large molecule, probably a protein or a nucleoprotein, which presumably penetrates reacting cells, though direct and unequivocal proof of such penetration is still unavailable. Inducing substances are active on vertebrates belonging to many different classes; *e.g.*, inductions of primary organs have been obtained by transplanting mammalian tissues into frog embryos or by transplanting tissues of a chick embryo into the embryo of a rabbit.

Induction is responsible not only for the subdivision of ectoderm into neural plate and epidermis but also for the development of a large number of organ rudiments in vertebrates. The notochord is a source of induction for the development of the adjoining somites and nephrotomes; the latter appear jointly to induce development of limb rudiments from the lateral plate mesoderm. Further examples are mentioned below in connection with development of the various organs.

Inducing substances

Since the results of induction are different for different organ rudiments, it must be presumed that there exist inducing substances with specific action, at least to a certain extent; thus, the lateral mesoderm induces differentiation of the skin but not neural plate from the very same kind of ectoderm. The number of inducing substances need not, however, be the same as the number of different kinds of tissues and organs, since certain differentiations could possibly be induced by a combination of two or more inducing substances, or the same inducing substance might have different effects on different tissues. It has been suggested that the regional organization of the entire vertebrate body could be controlled by the graded distribution of only two inducing substances—provisionally named the neuralizing substance and the mesodermalizing substance—along the length of the embryo. The neuralizing substance, concentrated at the anterior end, gradually decreases toward the posterior end; the mesodermalizing substance, on the

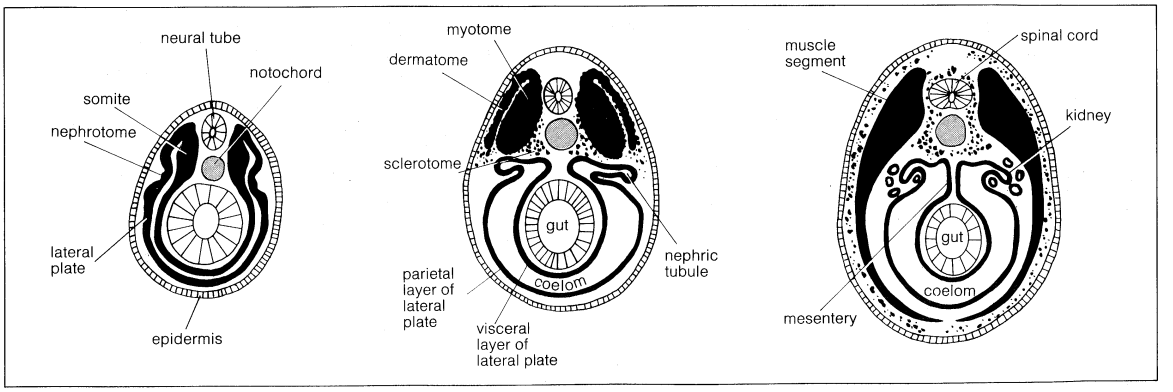


Figure 22: (Left to right) Subdivision of the mesoderm in a vertebrate embryo.

other hand, is concentrated at the posterior end and decreases toward the anterior end. The differentiation of induced structures depends on the relative amounts of the two inducing substances at any given point in the embryo. Acting alone, the neuralizing substance induces only nervous tissue, which takes the form of the forebrain, and the mesodermalizing substance induces only mesodermal structures (*e.g.*, somites, notochord).

In the amphibian embryo, induction appears to have its primary source in the dorsal lip of the blastopore, which eventually gives rise to the notochord and adjoining somites. Induction by the notochord and somites is responsible for the development of the neural plate in the ectoderm, of lateral and ventral parts of the mesodermal mantle, and of the lumen of the alimentary canal in the endoderm. The dorsal lip of the blastopore for this reason has been called the primary organizer. In higher vertebrates, in which gastrulation occurs through the medium of a primitive streak, the anterior end of the streak and the Hensen's node have properties similar to those of a primary organizer. Organization centres have been found, or suspected, in embryos of animals belonging to a few other groups, in particular the insects and sea urchins, but the interpretation of the experimental results in these animals is less satisfactory than in the case of vertebrates.

Concept of organization centre

The concept of an organization centre suggests that a part of the embryo differs from the rest of the embryonic tissues in being more active. The more active parts of the embryo (and also of animals in later stages of development) are particularly sensitive to certain noxious influences in their environment. If an embryo is deprived of oxygen or subjected to weak concentrations of poisons, the first parts to suffer are the most morphogenetically active ones. In vertebrate embryos the anterior end of the head is most sensitive. Early sea-urchin embryos have two centres of maximal sensitivity: one at the animal pole and the other at the vegetal pole. The damage done by noxious influences may result in actual breakdown of cells in a region of maximal sensitivity and may also lead to a depression of the developmental potential of the cells. Thus, the graded distribution of certain physiological properties appears to play a part in morphogenetic processes: physiological gradients are in fact also morphogenetic gradients.

Gradients in the embryo can be used to control development to a certain extent, by exposing the embryo to influences that, while reaching all parts, have a local effect as the result of differences in sensitivity. Disturbances of normal development often are the result of disruptions of gradients (see above *Aberrations of growth: biological malformation.*).

ORGANOGENESIS AND HISTOGENESIS

The primary organ rudiments continue to give rise to the rudiments of the various organs of the fully developed animal in a process called organogenesis. The formation of organs, even those of diverse function, shares some common features, which are considered in this section. As the organs form, so do their component tissues, in a process termed histogenesis.

A germinal layer, as the name implies, is a sheet of cells. An organ rudiment may be formed and separated from such a sheet in several ways. A groove, or fold, may appear within the layer, become closed into a tube, and then separated from the original layer. A tube once formed may be subdivided into sections by constrictions and dilations of the tube at certain points. This is the way the nervous system rudiment is formed in vertebrates as already described.

Alternatively, the germinal layer may produce a round depression, or pocket. The pocket may then separate from the layer as a vesicle, or it may elongate and branch at the tip while still connected with the layer. The latter method is common in the development of various glands and also the lungs in vertebrates.

Still another method of rudiment formation in a germinal layer is by the development of local thickenings, elongated or round, and detachment from the epithelial sheet. If a lumen appears later within such a body, the result may be the same as that achieved by folding—that is, a tube or vesicle may be formed. Indeed, the same sort of organ may develop even in related animals in either of these ways. The epithelial layer may further be cut up into segments, with the layer losing continuity, as in the formation of somites in vertebrates or similar mesodermal blocks in segmented invertebrates (*e.g.*, annelids and arthropods).

Embryonic mesenchyme

Lastly, the cells of a germinal layer may give up their connection to each other and become a mass of loose, freely moving cells called embryonic mesenchyme. This mass gives rise to various forms of connective tissue but may also condense into more solid structures, including parts of the skeleton and the muscles.

Many organs are comprised of all three germinal layers. It is very common for glands, for instance, to derive their lining from an ectodermal or endodermal epithelium and their connective tissue (sometimes in the form of a capsule) from mesenchyme of mesodermal origin. Parts of ectoderm and endoderm cooperate also in the development of the lining of the alimentary canal, and mesoderm provides the connective tissue and muscular sheath of the canal.

In this section the development of organs of the body are dealt with according to the germinal layer that contributes the most important part, and only the development of vertebrate organs is considered.

Ectodermal derivatives

THE NERVOUS SYSTEM

The vertebrate nervous system develops from the neural plate—a thickened dorsal portion of the ectoderm—which forms a tube, as described earlier. From the very start the tube is wider anteriorly, the end that gives rise to the brain. The posterior part of the neural tube, which gives rise to the spinal cord, is narrower and stretches as the embryo lengthens. Stretching involves the head to only a very minor degree.

The brain and spinal cord. Constrictions soon appear in the brain region of the neural tube, subdividing it into three parts, or brain vesicles, which undergo further transformations in the course of development. The most anterior of the primary brain vesicles, called the prosencephalon, gives rise to parts of the brain and the eye rudiments. The latter appear in a very early stage of development as lateral protrusions from the wall of the neural tube (Figure 21), which are constricted off from the remainder of the brain rudiment as the optic vesicles. The rest of the prosencephalon constricts further into two portions, an anterior one, or telencephalon, and a posterior one, or diencephalon. The telencephalon gives rise, in lower vertebrates, to the smell, or olfactory, centre; in higher vertebrates and man, it becomes the centre of mental activities. The diencephalon, with which the eye vesicles are connected, was presumably originally an optic centre, but it has acquired, in the course of evolution, a function of hormonal regulation. The floor of the diencephalon forms a funnel-shaped depression, the infundibulum, which becomes connected with the pituitary, or hypophysis, the most important gland of internal secretion (*i.e.*, endocrine gland) in vertebrates. Indeed, the posterior lobe of the hypophysis is actually derived from the floor of the diencephalon. Tissues of the infundibulum and the posterior lobe of the hypophysis produce certain hormones (oxytocin and vasopressin) and stimulate the production and release of other hormones from the anterior lobe of the hypophysis.

The second primary brain vesicle, the mesencephalon, gives rise to the midbrain, which, in higher vertebrates, takes part in coordinating visual and auditory stimuli.

The third primary brain vesicle, the rhombencephalon, is more elongated than the first two; it produces the metencephalon, which gives rise to the cerebellum with its hemispheres, and the myelencephalon, which becomes the medulla oblongata. The cerebellum acts as a balance and coordinating centre, and the medulla controls functions such as respiratory movements.

The cells constituting the wall of the neural tube and, later, of the brain and spinal cord become arranged in such a way that they point into the central cavity of the

Differentiation of nervous tissue

tube. The differentiation of nervous tissue involves many cells abandoning their connection to the inner surface of the neural tube and migrating outward, where they accumulate as a mantle. The first cells to migrate become the neurons, or nerve cells. They produce outgrowths called axons and dendrites, by which the cells of the nervous system establish communication with one another to form a functional network. Some of the outgrowths extend beyond the confines of the brain and spinal cord as components of nerves; they establish contact with peripheral organs, which thus fall under the control of the nervous system. Cells migrating from the inner surface of the neural tube later in development become astrocytes, which are the supporting elements of nerve tissue.

The fate of nerve cells is dependent largely on whether they succeed, directly or indirectly (through other neurons), in connecting with peripheral organs. Nerve cells that fail to establish connections die. Thus, if in early stages of embryonic development, some organ, a limb rudiment for instance, is surgically removed, the nerve cells in the centres supplying nerves to such an organ are reduced in number, and the corresponding nerves also diminish or disappear. On the other hand, if an organ is introduced by transplantation into a developing embryo, the organ will be supplied by nerves from a nerve centre in which the number of cells apparently increases; no additional cells are provided, but cells that would otherwise have degenerated remain active and differentiate into functional neurons, thus satisfying the demand created by the additional organ.

Nerves do not consist entirely of outgrowths of neurons located in the brain and spinal cord. Many components of nerves are outgrowths of neurons, the cell bodies of which are located in masses called ganglia; there are three main types of ganglia: spinal ganglia, cranial ganglia, and ganglia of the autonomous nervous system. The spinal ganglia are derived from cells of the neural crest—the loose mesenchyme-like tissue that remains between the neural tube and skin after separation of the two. Part of the cells of the neural crest in the region of the trunk and tail accumulate in segmental groups (corresponding to the mesodermal somites) and provide fibres to peripheral organs and to the spinal cord. These fibres constitute the sensory pathways in the spinal nerves. The motor components of the spinal nerves—fibres that activate muscles—are outgrowths of neurons lying in the spinal cord. The ganglia of the cranial nerves are produced only in part from cells of the neural crest; an additional component comes from the epidermis on the side of the head. Cells of the epidermal thickenings called placodes detach themselves and contribute to the formation of the cranial ganglia and thus of the cranial nerves.

The ganglia of the autonomous (sympathetic) nervous system are derived, as are the spinal ganglia, from neural-crest cells, but, in this case, the cells migrate downward to form groups near the dorsal aorta, near the intestine, and even in the intestinal wall itself. The outgrowths of cells in these ganglia are the nerve fibres of the sympathetic nerves (see also NERVES AND NERVOUS SYSTEMS).

Major sense organs. *The eye.* As has been pointed out, the rudiments of the eyes develop from optic vesicles, each of which remains connected to the brain by an eye stalk, which later serves as the pathway for the optic nerve. The optic vesicles extend laterally until they reach the skin, whereupon the outer surface caves in so that the vesicle becomes a double-walled optic cup. The thick inner layer of the optic cup gives rise to the sensory retina of the eye; the thinner outer layer becomes the pigment coat of the retina. The opening of the optic cup, wide at first, gradually becomes constricted to form the pupil, and the edges of the cup surrounding the pupil differentiate as the iris. The refractive system of the eye and, in particular, the lens of the eye are derived not from the cup but from the epidermis overlying the eye rudiment. When the optic vesicle touches the epidermis and caves in to produce the optic cup, the epidermis opposite the opening thickens and produces a spherical lens rudiment. The lens develops by an induction by the optic vesicle on the epidermis with which it comes in contact. A further influence emanating from the eye changes the epidermis remaining in place over the lens into a transparent area, the cornea. Influence of the optic cup on the surrounding mesenchyme causes the latter to produce a vascular layer around the retina and, outside of that, a tough fibrous or (in some animals) even a partly bony capsule called the sclera. Thus a complex interdependence of different materials produces the fully developed and functional vertebrate eye (see also SENSORY RECEPTION: *Photoreception*).

Formation of the optic cup

The ear. The main part of the ear rudiment is derived from thickened epidermis adjoining the medulla. This area of the epidermis invaginates to produce the ear vesicle, which separates from the epidermis but remains closely apposed to the medulla. The ear vesicle becomes complexly folded to produce the labyrinth of the ear. Subsequently, a group of cells of the ear vesicle becomes detached and gives rise to the acoustic ganglion. Neurons of this ganglion become connected by their nerve fibres to the sensory cells in the labyrinth, on the one hand, and with the brain (the medulla), on the other. The ear vesicle, acting on the surrounding mesenchyme, induces the latter to aggregate around the labyrinth and form the ear capsule. Further parts with various origins are added to the ear: the middle ear, from a pharyngeal pouch and the associated skeleton, and the external ear (where present), from epidermis and dermis.

The olfactory organ. The olfactory organ develops from a thickening of the epidermis adjacent to the neural fold at the anterior end of the neural plate. This thickening is converted into a pocket or sac but does not lose connection with the exterior. The openings of the sac become the external nares, and the cavity of the sac becomes the nasal cavity. Some cells of the olfactory sac differentiate as sensory epithelium and produce nerve fibres entering the forebrain. In most fishes the olfactory sac does not communicate with the oral cavity; in lungfishes and in terrestrial vertebrates, however, canals develop from the olfactory sacs to the oral cavity, where they open by internal nares. A cartilaginous capsule forms around the olfactory organ from cells believed to have been derived from the walls of the sac itself, and thus it is ectodermal in origin.

The nares and nasal cavity

Gustatory and other organs. Gustatory organs in the form of taste buds develop as local differentiations of the lining of the oral cavity but also, in fishes, in the skin epidermis. They are supplied with nerve endings, as are several other sensory bodies scattered among the tissues and organs of the developing body.

THE EPIDERMIS AND ITS OUTGROWTHS

The major part of the ectodermal epithelium covering the body gives rise to the epidermis of the skin. In fishes and aquatic larvae of amphibians, the many-layered epidermis is provided with unicellular mucous glands. In terrestrial vertebrates, however, the epidermis becomes keratinized; *i.e.*, the outer layers of cells produce keratin, a protein that is hardened and is impermeable to water. During the process of keratinization, many cell components degenerate and the cells die; the layer of keratinized cells is therefore shed from time to time. In reptiles the shedding may take the form of a molt in which the animal literally crawls out of its own skin. It is less well known that frogs and toads also molt, shedding the surface keratinized layer of their skin (which is usually eaten by the animal). In birds and mammals, keratinized cells are shed in pieces that are sloughed off, rather than in extensive layers. In many vertebrates local thickenings of the keratinized layer appear in the form of claws, hooves, nails, and horns.

Keratinization

The epidermis is only the superficial layer of the skin, which is reinforced by the dermis, a connective tissue layer of a much greater thickness. The cells of the dermis are derived from mesoderm and neural-crest cells. In particular the pigment cells found in the dermis of fishes, amphibians, and reptiles are of neural-crest origin. The pigment in the skin of birds and mammals (and also in hairs and feathers) is also produced by neural-crest cells, but in these animals the pigment cells penetrate into the epidermis or deposit their pigment granules there.

The structure of the skin is further complicated by the

development of hairs and feathers, on the one hand, and of skin glands, on the other. Hairs and feathers develop from a somewhat similar kind of rudiment. The development starts with a local thickening of the epidermal layer, beneath which a group of mesenchyme cells accumulate. In the case of hairs, the epidermal thickening proliferates downward and forms the root of the hair, from which the shaft then grows outward, emerging on the surface of the skin. In the case of feathers, the epidermal thickening bulges outward to form a hollow fingerlike protrusion with a connective tissue core. Secondarily, the shaft of the feather branches characteristically to produce barbs and barbules. In both cases, however, the final structure—shaft of the hair and shaft barbs and barbules of the feather—consists of keratinized and, thus, dead cells.

The skin of amphibians and mammals (but not of birds and reptiles) is provided with numerous skin glands, which develop as ingrowths from the epidermis. A peculiar type of skin gland is the mammary gland of placental mammals. In the first stage of development, mammary-gland rudiments resemble hair rudiments; they are thickenings of the epidermis, with condensed mesenchyme on their inner surfaces. In some mammals (rabbit, man) two continuous epidermal thickenings called mammary lines stretch along either side of the belly of the embryo. Parts of the line corresponding in number and position to the future glands enlarge while the rest of the thickening disappears. The initial thickenings proliferate inward and produce a system of ramified cords, solid at first but hollowed out later, which become the lactiferous, or milk-bearing, ducts of the gland. Further branching at the tips of the ducts gives rise to smaller ducts and to the secretory end sacs, or alveoli, of the gland.

Mesodermal derivatives

THE BODY MUSCLES AND AXIAL SKELETON

Derivatives of somites

The somites, formed in the early stages of development from the upper edges of the mesodermal mantle adjoining the notochord, are complex rudiments that subdivide and give rise to very diverse body structures. The coelomic cavity, present initially, becomes obliterated by the side-to-side flattening of the somites, so that the thinner, outer parietal layer of the somite comes in close contact with its thicker visceral layer. The visceral layer of the somite very early subdivides into two parts. The upper, dorsolateral part called the myotome (Figure 22) remains compact, giving rise to the body muscles. The lower, medioventral part of the somite, called the sclerotome, breaks up into mesenchyme, which contributes to the axial skeleton of the embryo—that is, the vertebral column, ribs, and much of the skull. The parietal layer of the somite, at a later stage, is converted into mesenchyme that, together with components of the neural crest, gives rise to the dermis of the skin and, for this reason, is called the dermatome.

The cells of the myotome are elongated in a longitudinal direction and become differentiated as muscle fibres. The myotomes, originally situated dorsally, expand on either side, penetrating between the skin on the outside and the lateral plates of the mesoderm on the inside, until they meet midventrally; the whole body is thus enclosed in a layer of developing muscle. As the somites and myotomes are segmented, so are the muscles derived from them. Metamerism, or segmentation, a feature in the embryos of all vertebrates, remains preserved only in the adults of fishes and of terrestrial vertebrates that have elongated bodies (salamanders, snakes); it becomes largely erased in four-footed animals that depend on their limbs for locomotion.

The mesenchyme derived from the sclerotomes condenses as cartilage around the notochord and the spinal cord. It forms the cartilaginous vertebral column and ribs. In the head region it produces a part of the cartilaginous skull, mainly its posterior and ventral parts; anteriorly the somitic mesenchyme is supplemented by mesenchyme from the neural crest. Cartilaginous capsules of the olfactory organ and the ear fuse with the cartilaginous capsule surrounding the brain; to this complex are also added cartilages associated with the jaws and gill skeleton. Cartilage in the vertebral column and in the skull is replaced later in the bony fishes and in the terrestrial vertebrates by bone. At a still later stage, dermal bones are added, which, while they have no precursors in the cartilaginous skeleton, develop in the adjoining mesenchyme.

THE APPENDAGES: TAIL AND LIMBS

The tail in vertebrates is a prolongation of the body beyond the anus. It develops in early stages from the tail bud, immediately dorsal to the blastopore. Material for the tip of the tail is situated slightly forward from the edge of the blastopore. The elongation of the back of the body is greater than that of the belly; as a result the tip of the tail bud is carried beyond the blastopore and thus beyond the anus, which, in the developed embryo, marks the position of the blastopore. The consequence is that a section of the dorsal surface of the embryo comes to lie on the ventral surface of the tail; *i.e.*, becomes inflected. The tail bud is formed from parts that have already been differentiated to a certain extent; prolongations of the neural tube and of the notochord are involved, and endoderm extends into the tail rudiment as the postanal gut, which, however, soon degenerates. The bud is also encased in ectodermal epidermis. In amphibians the somites of the tail are not derived from the chordamesodermal mantle but from the inflected posterior portion of the neural plate, which loses its nervous nature and becomes subdivided into segments corresponding to the somites of the trunk. In higher vertebrates the cells in the interior of the tail bud have an undifferentiated appearance and form a growth zone, at the expense of which parts of the tail (neural tube, notochord, somites) are extended backward as the tail elongates.

Formation of the tail bud

The paired limbs of vertebrates derive their first rudiments from the upper edge of the lateral plate mesoderm. The parietal layer becomes thickened, and cells escape from the epithelial arrangement and form a mesenchymal mass adjoining the ectodermal epithelium at the surface of the body. The ectodermal epithelium over the mass of mesenchyme likewise becomes thickened. In higher vertebrates, the accumulation of mesodermal cells and the thickening of the epidermis occur along the entire length of the trunk, from neck to anus, but in the middle of the trunk they soon disappear, and only the most anterior and the most posterior sections develop further into the rudiments of the forelimbs and hindlimbs, respectively. In fishes, the rudiments of the pectoral and pelvic fins are more extended anteroposteriorly in earlier than in final stages.

The mesodermal masses of the limb rudiments proliferate, and, covered with thickened epidermis, form on the surface of the body conical protrusions called the limb buds, which, once formed, possess all the materials necessary for limb development. Limb buds may be transplanted into various positions on the body or on the head and there develop into clearly recognizable limbs, conforming to their origin, whether a forelimb or hindlimb, a wing or a leg in birds. This specificity of the limb is carried by the mesodermal part of the rudiment, but a complex interaction between the mesodermal mesenchyme and the ectodermal epidermis is necessary for the normal development of the limb. In four-limbed vertebrates (tetrapods), the tips of the limb buds become flattened and broadened into hand or foot plates. The edge of the plate is indented, forming the rudiments of the digits. Meanwhile, local areas of the mesodermal mesenchyme in the interior of the limb rudiment condense; these are the rudiments of the various components of the limb skeleton. In fishes, small outgrowths from the myotomes enter the limb rudiment to form the muscles of the fins. In tetrapods, however, the limb muscles develop from the same mass of mesenchyme that gives rise to the skeleton. Thus the muscles of the body and the muscles of the limbs have different origins—the first develop from the myotomes (thus from the somites), and the second develop from the lateral plate mesoderm via the limb buds.

Limb buds

The nerves supplying the limbs grow into the limb rudiments from the spinal cord and the spinal ganglia. The nerves are guided in some way by the limb rudiments, for, if limb rudiments are displaced by transplantation to

an abnormal position, the nerves still find their way and establish normal relationships to the limb muscles. Limb rudiments transplanted to sites very far from their normal positions induce local nerves to enter the limb, thereby making it motile.

EXCRETORY ORGANS

Nephrotome development

The kidneys of vertebrates consist of a mass of tubules that develop from the stalks of somites called nephrotomes. In some primitive vertebrates such as cyclostomes, the nephrotome in each segment gives rise to only one tubule (Figure 22), but, in the great majority of vertebrates, mesenchyme from adjacent nephrotomes fuses into a common mass that differentiates into a number of nephric tubules irrespective of the original segmentation of the mesoderm. Under primitive conditions each tubule opens by a funnel (the nephrostome) into the coelomic cavity; the opposite ends of the tubules fuse to form the collecting ducts of the kidney. A collection of capillaries (the glomerulus) becomes associated with the nephric tubule, forming its filtration apparatus. The glomerulus may be situated in the coelomic cavity opposite the nephrostome or, in all the more advanced animals, intercalated into the nephric tubule, forming with the latter a renal corpuscle of the kidney. In adults of all vertebrates above the amphibians, the nephrostomes disappear (or are never formed), so that the tubule begins with the renal corpuscle. Parts of the kidney in vertebrates can be distinguished as the pronephros (most anteriorly, at the forelimb level), the mesonephros (in the midtrunk region), and the metanephros (in the pelvic region). The three sections of the kidney develop at different stages, starting with the pronephros and ending with the metanephros. In their morphology and mode of development, the anterior parts show more primitive conditions than the posterior ones. The pronephros, developing early in embryo formation, is the functional kidney of fish and amphibian larvae. Its collecting duct opens into the hindmost part of the intestine, called the cloaca, and later also serves as the collecting duct of the mesonephros. In reptiles, birds, and mammals, the pronephros is nonfunctional, although even in these animals its duct persists as the mesonephric duct. The mesonephros develops later and replaces the pronephros as the functional kidney of adult fishes and amphibians and of the embryos of reptiles, birds, and mammals. The tubules of the mesonephros link up with the duct derived from the pronephros. The pronephric duct in fact stimulates the development of mesonephric tubules, and, in its absence, the mesonephros does not develop at all.

Development of the metanephros

The metanephros is found only in reptiles, birds, and mammals. It replaces the mesonephros of the early embryonic stages and continues as the functional kidney in the postembryonic and adult life of these animals. The metanephros develops from mesenchyme derived from the nephrotomes of the posterior part of the trunk and lying dorsal to the mesonephric duct. The actual differentiation is initiated by a dorsal outgrowth of the mesonephric duct, called the ureteric bud. The ureteric bud grows in the direction of the mesenchyme and becomes the ureter. Having penetrated the mass of mesenchyme, it starts to branch, producing the collecting tubules of the kidney; the mesenchyme, meanwhile, in response to the influence of the duct and its branches, aggregates to form the excretory tubules of the kidney. The influence of the ureter is indispensable for the development of the metanephric excretory tubules, for, if the ureter fails to develop or, in its outgrowth, stops short of reaching the kidney-producing mesenchyme, no kidney develops.

CIRCULATORY ORGANS

The rudiment of the heart in vertebrates develops from the ventral edges of the mesodermal mantle in the anterior part of the body, immediately adjoining the pharyngeal region. A group of mesodermal cells breaks away from the ventral edge of the lateral plate, takes a position just underneath the pharyngeal endoderm, and becomes arranged in the form of a thin-walled tube, which will become the endocardium, or lining of the heart (Figure 23). In vertebrates with complete cleavage, the endocardial tube is single and

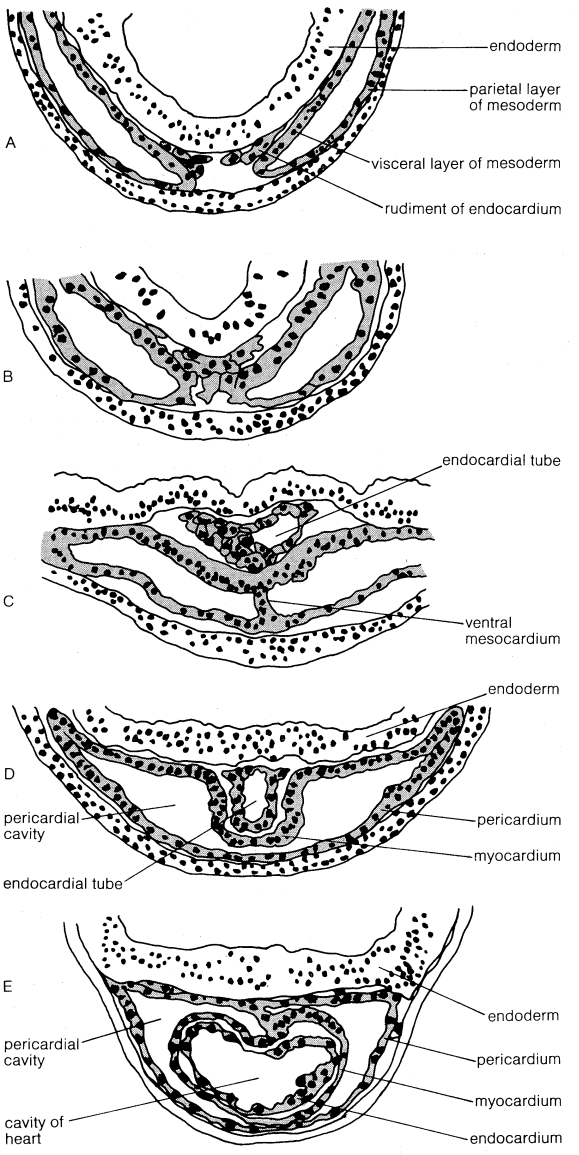


Figure 23: Development of the heart in amphibian embryos.
From O. Hertwig, *Handbuch der Vergleichendenden und Experimentellen Entwicklungslehre der Wirbeltiere* (1906); Gustav Fischer Verlag

medial from its start. In higher vertebrates with meroblastic cleavage—reptiles, birds, and mammals—the embryo in early stages of development is flattened out on the surface of the yolk sac; therefore, what are morphologically the ventral edges of the mesodermal mantle lie far apart on the perimeter of the blastodisc. As a result of this arrangement, two endocardial tubes are formed, one on either side of the embryo. Subsequently, when the embryo becomes separated from the yolk sac, the two endocardial tubes meet in the midline ventral to the pharynx and fuse, producing a single heart rudiment. After the formation of the endocardium, or the lining of the heart, the coelomic cavity in the lateral plate mesoderm adjoining the heart rudiment expands slightly and envelops the endocardial tube or tubes (Figure 23). The heart muscle layer, or myocardium, develops from the visceral (splanchnic) layer of the lateral plate that is in contact with the endocardial tube; the parietal (somatic) layer of the lateral plate forms the pericardium, or covering of the heart. The portion of the coelom surrounding the heart becomes separated from the rest of the body cavity and develops into the pericardial cavity.

The endocardial tube branches anteriorly into two tubes, the ventral aortas; a similar branching of the endocardial tube posteriorly forms the two vitelline veins, which carry blood from the midgut endoderm or from the yolk sac (when present) to the heart.

In its earliest development, the heart rudiment shows a

degree of dependence on the adjoining endoderm. The whole of the endoderm can be removed in newt embryos in the neural-tube stage. In such endodermless embryos, the heart fails to develop, even though the mesoderm destined to form the heart rudiment is left intact.

Development of the heart

The heart is initially a straight tube stretching in an anteroposterior direction. Rather early in development, however, it becomes twisted in a characteristic way and subdivided into four main parts: the most posterior, the sinus venosus; the atrium, which comes to lie at the anteriorly directed bend of the tube; the ventricle, occupying the apex of the posteroventrally directed inflexion; and, most anteriorly, the conus arteriosus. In the course of development in the more advanced vertebrates, the atrium and ventricle become partially or completely subdivided into right and left halves. In amphibians, only the atrium is separated into two halves, by a partition starting from the posterior end. In reptiles, a partition separates the atria and part of the ventricle. In birds and mammals, the subdivision of the heart is complete, with two atria and two ventricles.

The complete subdivision of the heart is important for separating the pulmonary, or lung, blood supply from the general body circulation. But, if this separation developed early in the embryo, it would create difficulties, since the lungs of the embryo are not functional; the enrichment of the blood with oxygen occurs instead in the placenta. The partition between the atria in mammalian embryos remains incomplete, so that blood returning from the body and from the placenta enters into the right half of the heart but is shunted (through the interatrial foramen) into the left half of the heart and thence again into general circulation. At birth, however, the interatrial foramen is closed by a membraneous flap, and oxygen-depleted blood from the body enters the right atrium, is channelled into the right ventricle, and thence to the lungs for oxygenating.

In an adult vertebrate, blood vessels extend to all parts of the body. It would seem that channels for the supply of blood are provided in proportion to the local demand of the tissues; progressively developing organs or parts with particularly intensified function always receive an increased blood supply. The rudiments of blood vessels are always aggregations of mesenchyme cells. In any blood vessel the endothelial tube is formed first, and the muscular and elastic layers are added later.

The main blood channels in vertebrates develop in certain favoured situations; namely (1) between the endoderm and lateral plate mesoderm; (2) around the kidneys, especially the pronephros and mesonephros; and (3) in connection with the heart, which is a special case of the first category.

The aortic arches

From the paired forward extensions from the heart, the ventral aortas, loops develop between the pharyngeal clefts. These are the aortic arches (Figure 24), which served originally to supply blood to the gills in aquatic vertebrates. The arches are laid down in all vertebrates, six or more being found in cyclostomes and fishes; six are present in the embryos of tetrapods, but the first two are degenerate. The arches of the third pair develop as the carotid arteries, supplying blood to the head. Those of the fourth pair (and, exceptionally, in urodeles also the fifth) join dorsally to form the dorsal aorta, providing blood to most of the body. These are the systemic arches. The arches of the sixth pair are the pulmonary arches; in embryos they carry blood to the dorsal aorta, as well as to the lungs, but in fully developed amniotes (reptiles, birds, and mammals), they carry blood only to the lungs.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

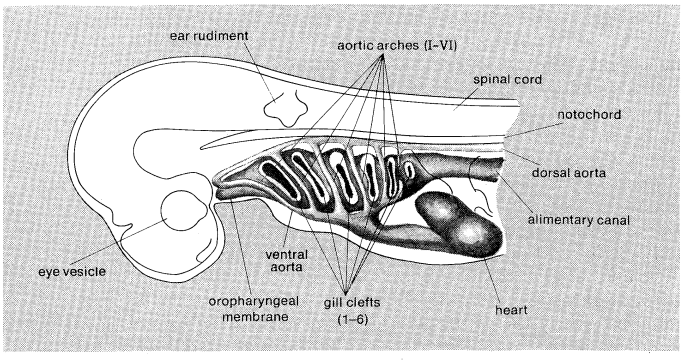


Figure 24: Relation of the aortic arches and the branchial clefts in a dogfish embryo.

The paired posterior extensions of the heart of the early embryo are the vitelline veins, whose branches spread out between the lateral plate mesoderm and the endoderm, especially the endoderm of the yolk sac, when present. On their way to the heart, the vitelline veins pass through the liver and break up into a system of small channels—the hepatic sinusoids. Parts of the vitelline veins lying posterior to the liver become the hepatic portal veins, which carry blood from the intestine to the liver; the parts of the vitelline veins anterior to the liver become the hepatic veins, which carry blood fom the liver to the sinus venosus in lower vertebrates (anamniotes), but become the anterior section of the postcaval vein in amniotes.

Whereas the vitelline veins and, later, the hepatic portal vein carry blood from the endodermal parts of the embryo and from the yolk sac to the heart, the blood from the mesodermal and ectodermal parts is returned to the heart through a system of cardinal veins (Figure 25). These latter veins start their development in the form of an irregular

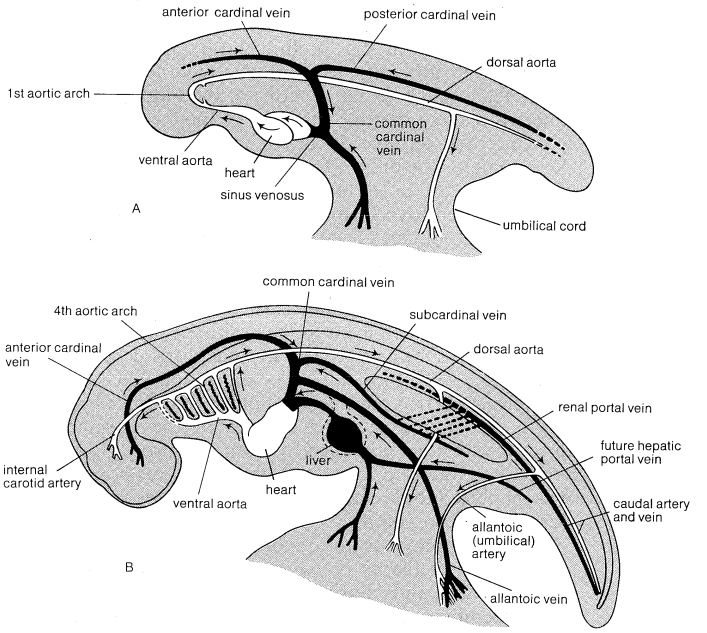


Figure 25: General arrangement of blood vessels in (A) an early amniote embryo and (B) the further development of the circulatory system. The arteries are shown in white, the veins in black.

Development of the cardinal veins

sinus around the pronephros, connected by the common cardinal veins (ducts of Cuvier), on either side, to the sinus venosus. Extensions anteriorly and posteriorly give rise to the precardinal and postcardinal veins, respectively. The postcaval vein, present in terrestrial vertebrates, is a late acquisition, both in evolution and in embryogenesis; it is a result of the intercommunication of several venous channels, including the anterior portion of the vitelline veins.

The first blood cells in vertebrate embryos form in association with the intestinal endoderm on the yolk sac. Groups of mesoderm cells derived from the splanchnic layer of the lateral plate (extra-embryonically in cases in which a yolk sac is present) become so-called blood islands, which are particularly conspicuous on the yolk sac of bird embryos (in the area vasculosa). In bird's eggs, the internal cells of the blood island start producing hemoglobin (gas-carrying component of blood) and become the first red blood cells (erythrocytes) as early as the second day of incubation. The outer cells of the blood islands develop into an endothelial layer and form a network of blood vessels covering part of the surface of the yolk sac. The network

acquires a connection to the vitelline veins and vitelline arteries (the latter being branches of the dorsal aorta); thus the blood corpuscles formed in the blood islands can enter the general blood circulation.

At later stages of embryogenesis, blood-cell formation shifts from the blood islands to the liver and, still later, to the bone marrow.

The lymphatic system, in a manner similar to the blood vessels, develops by the local aggregation of connective tissue to form lymphatic vessels.

REPRODUCTIVE ORGANS

In considering the development of reproductive organs, distinctions must be made between: (1) the origin of sex cells (gametes), (2) the origin and differentiation of the sex glands, or gonads (ovaries and testes), and (3) the origin and development of the supporting parts of the reproductive system (*e.g.,* genital ducts, copulatory organs).

Origin of germ cells

The germ (germinal) cells, which eventually give rise to the gametes, are often segregated from the somatic, or body, cells at a very early stage—during cleavage and before the subdivision of the embryo into ectoderm, mesoderm, and endoderm. In the invertebrate nematodes, the very first of these primordial germ cells is identifiable after as few as five divisions of the egg cell. The germ cell retains the large chromosomes present in the fertilized egg; in the somatic cells the chromosomes become fragmented. Subsequently, the single germ cell gives rise, by mitotic divisions, to all the gametes in the gonad.

In vertebrates, primordial germ cells arise outside the gonads, but they cannot be distinguished in early cleavage stages. In amphibians, cytoplasm at the vegetal pole, rich in ribonucleic acids, becomes incorporated into a number of cells, which, during cleavage and gastrulation, lie among the yolky endoderm cells. Later they migrate into the mesodermal layer and become incorporated into the rudiments of the gonads. In higher vertebrates, primordial germ cells can be recognized in the extra-embryonic endoderm of the yolk sac. In mammals, these cells subsequently migrate into the mesoderm and are located in the gonad rudiments. The mouse embryo, for example, originally has fewer than 100 primary germ cells; during their migration, however, their numbers increase as a result of repeated divisions, to 5,000 or more in the gonads.

Although the primordial germ cells either may appear before the separation of germinal layers or be found originally in the endoderm, the gonads are invariably of mesodermal origin. In vertebrates, the first trace of gonad development is a thickening of the coelomic lining on either side of the dorsal mesentery and medial to the kidney rudiments. The thickening, elongated anteroposteriorly, is known as the germinal ridge (Figure 26). The ridge protrudes into the coelomic cavity, and the fold of thickened epithelium becomes filled with mesenchyme. At this stage the primordial germ cells invade the rudiments of the gonads and become associated with the somatic cells of the germinal ridge. In the functionally differentiated gonads, only the actual gametes and their predecessors (spermatogonia and oogonia) are derived from the primary germ cells; the supporting cells are somatic cells of local mesodermal origin. In the ovaries, the follicle cells surrounding and nourishing the young egg cells (oocytes) are of somatic origin, as are also the connective tissue and blood vessels of the gonad. In the testes, supporting elements called Sertoli cells are somatic cells, as are the interstitial cells, which are scattered between the sperm-carrying tubules of the testes and believed to be the source of male hormones.

In the early stages of their development—even while the gonad rudiment is being invaded by primordial germ cells—the female and male gonads are in an indifferent stage. Only later does tissue differentiation of the gonads begin and male or female gonadal development proceed.

Development of genital ducts

The genital ducts, by which the eggs and sperm are carried away from the gonads, are, in vertebrates, linked with the excretory system. In the male, the seminiferous tubules connect with the nephric tubules of the mesonephros, and the sperm are carried to the exterior by way of the mesonephric duct. In males of lower vertebrates, the mesonephric duct thus serves as a channel

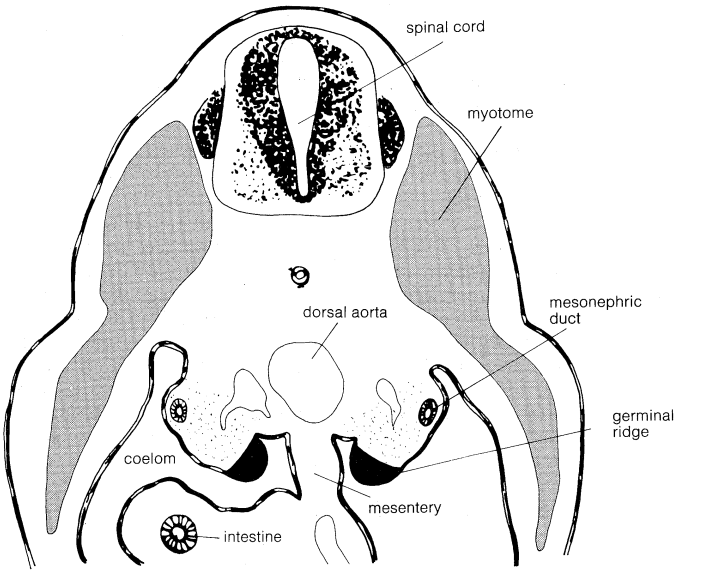


Figure 26: Transverse section of a mouse embryo showing the position of the germinal ridges and their relation to the mesonephric rudiment and the dorsal mesentery.

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

both for urine and for sex cells. In amniotes the development of the metanephros as the urine excreting organ has freed the mesonephric duct to carry products associated only with reproduction. In the female, a separate duct, the paramesonephric duct (Müllerian duct), develops beside the mesonephric duct. At its anterior end it utilizes the funnels of the pronephric tubules as its entrance (ostium). The paramesonephric duct develops initially in both female and male embryos. The ducts remain in an indifferent stage longer than the gonads. Eventually the sex hormones produced by the differentiating gonads cause a corresponding differentiation of the ducts. The mesonephric ducts, which become reduced in female embryos, remain in male embryos as ducts for conveying sperm (ductus deferens). The paramesonephric ducts, on the other hand, degenerate in male embryos but become the oviducts in female embryos (Figure 27). In mammals, the terminal portions of the paired oviducts differentiate as two uteri, which, in primates and man, fuse to form a single uterus.

In all terrestrial vertebrates except the placental mammals, the genital ducts, as well as the ducts of excretory organs, open into the cloaca. In mammals, however, the

From B.I. Balinsky, *An Introduction to Embryology*, 3rd ed. (1970); W.B. Saunders Company, Philadelphia

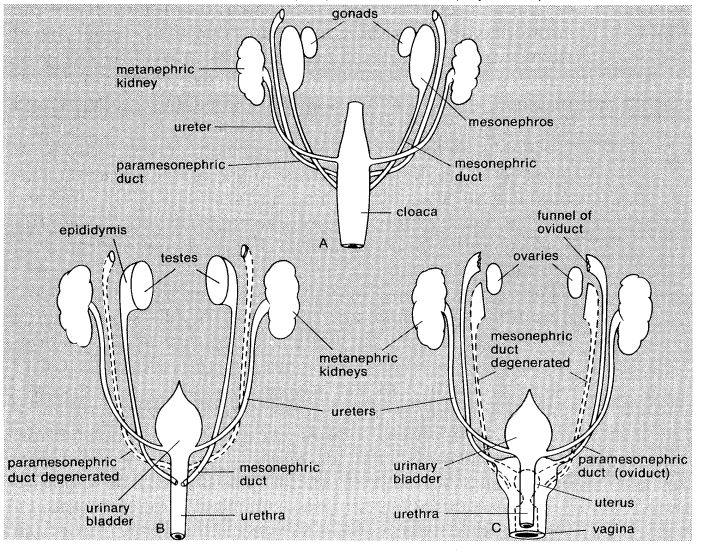


Figure 27: Transformations of the genital ducts in mammalian embryos in transition from an (A) indifferent stage to (B) the male and (C) female condition.

cloaca becomes subdivided into a dorsal part, which conveys the feces, and a ventral part, which receives excretory and genital products. In male mammals the excretory and genital ducts remain connected, having the urethra as their common outlet; in females the urethra serves only for the passage of urine and the uterus opens separately by means of the vagina. In nearly all vertebrates, the male nephric duct is utilized in some degree for the conduction of sperm.

Copulatory organs have developed independently in several groups of vertebrates having internal fertilization. The penis in mammals develops from an outgrowth called the genital tubercle, located at the anterior edge of the urinogenital orifice. The tubercle is laid down in a similar way in embryos of both sexes, and the region of the urinogenital orifice remains in an indifferent state even longer than do the genital ducts. In a comparatively late stage of embryonic life the genital tubercle of male embryos encloses the urethral canal and becomes the penis; in female embryos it remains small and becomes the clitoris.

Endodermal derivatives

THE ALIMENTARY CANAL

The alimentary canal is the chief organ developing from endoderm. The way it forms depends on the type of egg cleavage. In eggs with holoblastic (complete) cleavage, after gastrulation the invaginated mass of endoderm lines the archenteron, the cavity of which becomes the alimentary canal, or gut. In eggs with meroblastic (partial) cleavage—and also in mammals (despite their complete cleavage)—the endoderm is produced in the form of a sheet lying flat over the yolk-sac cavity. Subsequently, folds of endoderm and splanchnic mesoderm appear—first anteriorly, then laterally, and lastly posteriorly—and sink, converging ventrally under the embryo and cutting off the future gut cavity from the cavity of the yolk sac. The most anterior and posterior portions of the gut separate, but the middle part remains in open communication with the yolk sac throughout embryonic life, eventually becoming reduced to the yolk stalk, which passes through the umbilical cord.

The alimentary canal of vertebrates becomes differentiated into the oral cavity, pharynx, esophagus, stomach, and intestine. Whether derived from an archenteron or formed by folding of the endodermal sheet, the canal initially does not possess an opening at its anterior end. This is also the case in some lower chordates and echinoderms, which are grouped together with vertebrates as the Deuterostomia, or animals with secondary mouths.

'ormation
f
ertebrate
nouth

In vertebrates, a mouth forms by a rupture at the anterior end, where the endoderm is in contact with ectoderm. The ectoderm of the future mouth region becomes depressed, forming a mouth invagination, or stomodaeum. The ectodermal and endodermal layers separating the cavity of the stomodaeum from the gut fuse to form the oropharyngeal membrane, which thins and ruptures, providing free passage from the exterior to the gut. Because of its mode of origin, the oral cavity is in part lined by ectoderm and in part by endoderm, the two parts becoming indistinguishable. Before the oropharyngeal membrane ruptures, however, a small pocket forms on the dorsal side of the stomodaeal invagination. This, the rudiment of the anterior lobe of the hypophysis, becomes apposed to the ventral surface of the diencephalon and loses its connection with the mouth cavity.

The anal opening in some exceptional cases (urodele amphibians) is derived directly from the blastopore, which persists as a narrow canal after completion of gastrulation. In other vertebrates, however, the anus develops either near the location of the former blastopore or in a corresponding region at the posterior end of the embryo, where the last remnants of mesoderm migrated to the interior. It is thus claimed that the anus in vertebrates is derived, directly or indirectly, from the blastopore. The mode of formation of the opening is somewhat similar to that of the mouth. A slight invagination of the ectoderm occurs, and a cloacal membrane forms, separating the ectodermal invagination from the gut cavity. The membrane ruptures later to provide the anus.

THE PHARYNX AND ITS OUTGROWTHS

The anterior portion of the endodermal gut, lying immediately posterior to the mouth cavity, expands laterally as the pharynx. The lateral pockets of the pharyngeal cavity, called the pharyngeal pouches, perforate the mesodermal layer, reach the ectoderm, and break through to form pharyngeal, or gill, clefts (Figure 24). In fishes and larvae of amphibians, these clefts develop gills and become respiratory organs. Pharyngeal pouches develop in the early embryos of all vertebrates, including the airbreathing terrestrial reptiles, birds, and mammals. The number of pouches has been reduced in the course of evolution from six or more to four in tetrapods, and the posterior pouches may not actually break through.

The consistent development of pharyngeal pouches and clefts indicates their importance in vertebrate development. Many parts of the vertebrate body are derived from, or dependent on, the pharyngeal pouches; for example, the aortic arches—the most important blood vessels of a vertebrate—develop between successive pharyngeal pouches (Figure 25). Skeletal visceral arches also occur between consecutive pharyngeal pouches (they do not develop if the pharyngeal pouches are prevented from developing). In adult terrestrial vertebrates, parts of the visceral arches are transformed into the hyoid apparatus, supporting the tongue, the auditory ossicles, and parts of the larynx and trachea. Furthermore, some of the material of the pharyngeal pouches is utilized for the formation of the parathyroid glands and the thymus; the former are indispensable glands of internal secretion, and the latter are a source, in mammals, of cells that produce antibodies. The pharynx also produces the rudiment of the thyroid gland as a ventral outgrowth.

THE LIVER, PANCREAS, AND LUNGS

Three additional important organs develop from the endoderm: the liver, the pancreas, and the lungs. The liver develops as a ventral outgrowth of the endodermal gut just posterior to the section that eventually will become the stomach. Initially, the liver takes the form of a tubular gland, but it soon acquires a close relationship to the blood sinuses and capillaries, forming lobules around blood vessels rather than around glandular ducts. The pancreas develops from three independent rudiments: two ventral ones, formed just posterior to the liver rudiment, and a dorsal one. The ventral and the dorsal rudiments fuse in most vertebrates to form one organ with a complicated system of ducts opening into the duodenum, a portion of the small intestine. The lungs develop from a ventral hollow outgrowth of the gut, which is located just posterior to the pharyngeal region; the outgrowth branches into a right and left trunk that grow posteriorly beside the esophagus and then expand into hollow sacs, in lower terrestrial vertebrates, or into a system of tubes, in birds and mammals.

Development of lungs

The endodermal parts of the alimentary system are, along their entire length, encased by the splanchnic mesoderm of the lateral plates. The coelomic cavities of the right and left sides fuse ventral to the gut but remain separated dorsally by their respective walls, which form the dorsal mesentery—a double membrane by which the gut is suspended from the dorsal side of the body cavity and through which blood vessels and nerves reach the gut (Figure 22). The layer of splanchnic mesoderm next to the endoderm produces the connective tissue and muscular layers of the gut. During development of the glands of the alimentary canal (*e.g.*, pancreas, salivary glands), the mesoderm forms a connective tissue capsule around the branching tubules of the gland. The development of the tubules is dependent on this mesodermal capsule and cannot proceed without it.

Postembryonic development

After partially developing within the egg membranes or within the maternal body, the newly formed individual emerges. The new animal is then born (ejected from the mother's body) or hatched from the egg. The condition of the new organism at the time of birth or hatching

differs in various groups of animals, and even among animals within a particular group. In sea urchins, for example, the embryo emerges soon after fertilization, in the blastula stage. Covered with cilia, the sea-urchin blastula swims in the water and proceeds with gastrulation. Frog embryos emerge from the egg membranes when the main organs have already begun to develop, but functional differentiation of the tissues is unfinished; for instance, the components of the eyes and ears are far from complete, the mouth is not yet open, and the gut is filled with yolk-laden cells. Certain birds (called precocial) emerge from the egg covered with downy feathers and can run about soon after hatching, whereas others (altricial) hatch naked, with only rudiments of feathers, and are quite unable to move around. Among mammals there is a great range in the degree of development at birth. In marsupials, such as opossums and kangaroos, the young are born incompletely developed and very small; the young are then kept for a long time in the pouch of the mother, all the while firmly attached to the teats and suckling. Many small mammals are helpless at birth. Mice are born naked and blind; puppies and kittens are born covered with fur but with unopened eyes. Newborn human babies have their eyes open but cannot move themselves about for several months. Hoofed mammals, on the other hand, bear young that can stand up and run after their mothers within a few hours of birth.

Egg tooth of birds

In birds the hard shell is broken by the hatchling's beak, which is provided with a sharp tubercle on its top. A similar "egg tooth" appears on the tip of the snout of hatchling reptiles. Many arthropods have a preformed line of fragility that allows part of the eggshell to be burst open like a lid, allowing the young to emerge. Birth in mammals is effected through the contraction of smooth muscles of the uterus.

THE LARVAL PHASE AND METAMORPHOSIS

The organism emerging from the egg or from the maternal body, apart from being incompletely developed, may have an organization more or less different from that of an adult. In some cases the difference is so great that, without knowing the origin of the eggs or without following the young through their full course of development, it would be impossible to know that the young and the adult are of the same animal species. Such young, called larvae, transform into the adult form by a process of metamorphosis. The term larva also applies to young that resemble the adult form but differ from it in some substantial respect, as in possessing organs not present in the adult or in lacking an important structure (apart from sex glands and associated parts, which tend to develop later in life in most animals). Larvae in different animals have special names given to them, such as the tadpole of frogs, the caterpillar of butterflies, and the fry of fishes.

Advantages of larvae

The development of the embryo into a larva rather than directly into an organism similar to the adult has various advantages. At the time of emergence from the egg, the new individual is relatively small, and the organization that enables the adult to lead a particular mode of life may not be suitable for a miniature copy of the adult. The larva may have to procure food for itself and, being small, may not be able to feed in the same way as the adult. It also may not be able to use effectively the same defense mechanisms the adult possesses. The larval stage enables an animal to avoid such hazards; it provides a mode of life and corresponding organization better suited to the smaller size of the newly emerged organism. Another advantage is that the larva may be able to exploit an entirely different environment because its organization is very different from that of the adults. A terrestrial adult may have aquatic larvae, a flying adult may have burrowing larvae, and a parasitic adult may have a free-living larva. A third advantage of a larval stage emerges in animals whose adult stages are sessile or restricted in their movements; the larvae can move freely, either of their own accord or on water currents. In this way the larvae of sedentary animals serve for the dispersal of the species. Lastly, the larval stage is of great advantage for certain internal parasites, which, once inside a host, cannot transfer to another. New hosts are infected instead by the larval stages. (The usual means of attaining this end is for the parasite to produce enormous quantities of eggs and rely on the passive entry of the eggs into the new host with food. A more efficient way, however, is for a mobile larva to enter the new host actively.)

A large number of marine invertebrates possess floating larvae that have hairlike projections (cilia) as their means of locomotion. There are three main types of larvae, characteristic of large subdivisions of the animal kingdom.

The planula larva of coelenterates has an elongated shape and cilia covering its entire surface. The internal organization is simple, hardly beyond differentiation into ectoderm and endoderm in the interior. The larva does not feed but serves only for dispersal.

The trochophore larvae

The trochophore larva is found in many marine invertebrates. Typically, as in polychaetes, it has an alimentary canal with mouth and anus and a ring of ciliated cells arranged anterior to the mouth. It also possesses a sensory organ and rudiments of mesoderm. Cilia around the mouth bring in food—unicellular plants and other small particles. The larva thus not only serves for dispersal but also feeds and grows before it transforms into an adult worm. Other trochophore larvae are found in marine mollusks and in certain marine worms. The larva of echinoderms is similar to the trochophore in possessing a gut and a ciliary band, but the arrangement of the latter is different. The echinoderm larva also feeds and grows as well as serves for dispersal.

Larvae of very different kinds are found in many arthropods. In crustaceans the larva, called nauplius, does not differ substantially in mode of life or means of locomotion from the adult but has fewer appendages than the adult. A typical crustacean nauplius has three pairs of legs and an unpaired simple eye. Additional pairs of appendages and paired compound eyes appear in the course of a sometimes prolonged development. In insects the larva differs from the adult by the absence of wings but, in addition, may have a different mode of life and different way of feeding. Among chordates the tunicates (sea squirts) deserve attention; the larval form is a free-swimming creature, showing unmistakable relation to vertebrates, but the adult is sedentary, with much reduced nervous and muscular systems. The tadpole of a frog differs from the adult in being totally aquatic, in possessing a tail and gills for respiration, and in having a mouth adapted for feeding on plants. The adult frog is adapted to land life, except for reproductive periods, has no tail and no gills, and is an active predator.

Types of metamorphosis

Metamorphosis, the transformation of the larva into an adult, is a more or less complicated process depending on the degree of difference between the two forms. The transformation may be gradual, extend over a long period, and involve a number of intermediate stages; alternatively, the transformation may be achieved in one step. In the latter case, especially if the difference between the larva and adult is great, large parts of the body of the larva, including all the specifically larval organs, disintegrate (necrobiotic metamorphosis). At the same time, organs of the adult are built up, sometimes from reserve groups of cells that remain undifferentiated or nonfunctional in the larva. A good illustration of the distinction between gradual and abrupt metamorphosis occurs among the insects. In more primitive insects, such as cockroaches and grasshoppers, metamorphosis is gradual. The larva, often referred to as a nymph, has more or less the same organization as the adult, or imago; it feeds in a similar way but differs from the adults in lacking wings and in having incomplete sex organs. The wings appear in later stages of larval life; they are small at first but increase with each molt, and they attain full size and functional capacity at the last (imaginal) one. The larva of other insects, such as beetles, butterflies, and wasps, is a grub or caterpillar, a wormlike creature not even remotely resembling the adult. The difference in organization is so profound that the transformation cannot be achieved gradually, and an intermediate resting, or pupal, stage is interposed between the larva and imago. The pupa neither feeds nor moves, as the larval organs inside are destroyed and replaced with organs of the adult,

The nymph stage

including wings and sex organs. Eventually, when formation of the adult organs is complete, the pupal skin is cast off, and the adult emerges. The destruction of the larval parts may be far reaching and include even the skin and most of the alimentary canal. The tissues of the adult are formed from groups of reserve cells that were present all along in the larva as imaginal disks.

Necrobiotic metamorphosis is observed in the tunicate larva, in which the tail, including notochord, nerve cord, and muscles, and most of the brain, including eye and statocyst, are destroyed at the same time that the large pharyngeal cavity of the adult develops. A tadpole metamorphosing into an adult frog loses its tail—the cells of which are destroyed and devoured by phagocytic cells—its gills, and its larval mouthparts; concurrently the legs of the adult frog develop progressively, the structure of the mouth and alimentary canal change, and the skin acquires a bony (keratinized) layer and a system of subcutaneous glands (Figure 28).

From E. Witschi, *Development of Vertebrates* (1956); W.B. Saunders Company, Philadelphia

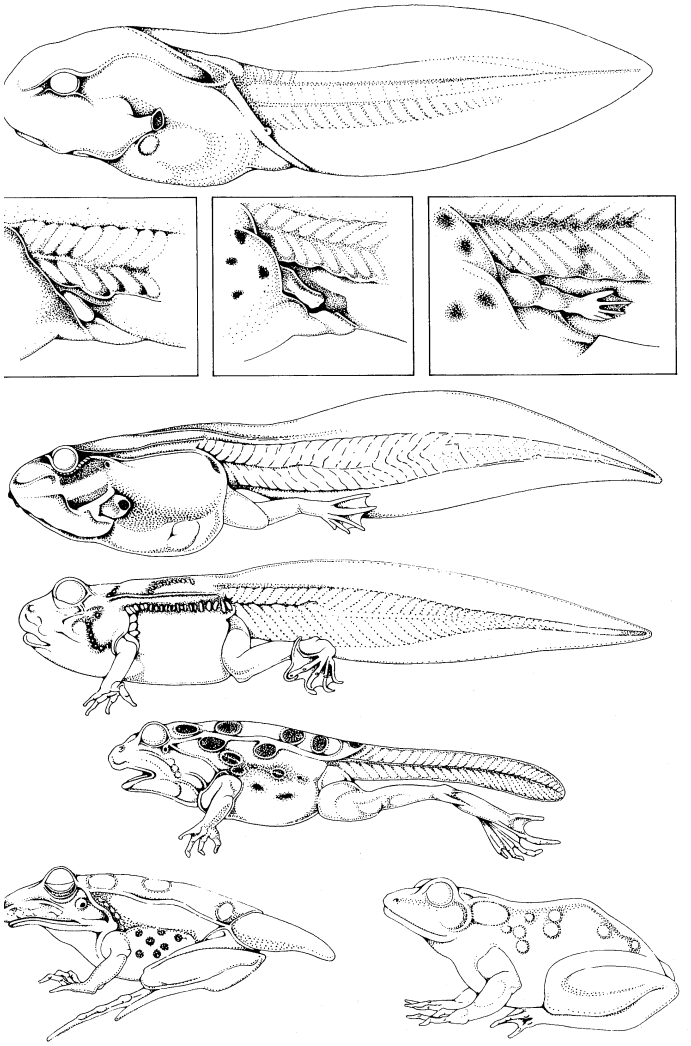


Figure 28: Stages in the metamorphosis of the frog from aquatic tadpole to terrestrial adult frog, including details of development of the hindleg.

The complicated changes taking place during metamorphosis, especially in the case of necrobiotic metamorphosis, must be performed in a coordinated way. So that no changes are made prematurely and no organ systems are left behind in the general transformation, some common signal for the change must be provided. For both insect and amphibian metamorphoses, which have been the most extensively studied, the signal is a hormonal one, sent in the blood to all the cells and tissues of the body.

Molting and hormonal function

Metamorphosis in an insect is complicated by the fact that the rigid cuticle covering its body is very restrictive; new features can appear only after a molt, when the old cuticle is replaced by a newly formed one. Molting in insects is caused by the action of two hormones. In the brain of insects, several groups of neurosecretory cells produce the first hormone. This brain hormone does not itself affect molting but stimulates the prothoracic gland, a loose mass of secretory cells situated in the thorax in close association with tracheal tubes. In response to the stimulation by the brain hormone, the prothoracic gland releases into the blood a second hormone, the molting hormone, or ecdysone. Under the influence of ecdysone, the tissues of the body produce a new cuticle under the old one, after which the old cuticle is shed (the actual molting). The new cuticle embodies any new developmental features that were scheduled to appear. The kind of feature that emerges after a molt is controlled by a third organ of internal secretion, the corpus allatum, secretory tissue situated posterior to the brain, near or around the dorsal aorta and usually appearing as a pair of separate or fused organs. The corpora allata emit the juvenile hormone, which, as long as it circulates in the blood, acts to perpetuate the larval form. As the larva approaches the end of its development, however, the corpora allata stop producing juvenile hormone or reduce its quantity; whereupon, the larva, at the next molt, metamorphoses into an adult. Withdrawal of the juvenile hormone is the immediate cause of metamorphosis, in conjunction with the brain hormone and ecdysone, which are responsible for the shedding of the larval cuticle and for the production of the new cuticle embodying the features of the imago. Metamorphosis through the stage of the pupa is effected by diminishing levels of juvenile hormone, which determine first the transformation of the larva into a pupa and, with further reduction of the juvenile-hormone level, the final step of transformation of the pupa into the adult.

The metamorphosis of a tadpole into a frog also depends upon two hormones: one initiating the process and the other directly influencing the tissues involved in the change. The first hormone is the thyrotropic hormone, produced by the hypophysis. It has no immediate effect on the tissues of the body but activates the thyroid gland to produce several substances, the most important of which is thyroxine. Thyroxine and other iodine-containing compounds circulate in the blood and cause changes that, in their entirety, constitute the process of metamorphosis. It is remarkable that different tissues react in different ways to the presence of thyroxine. The muscles of the tadpole's tail degenerate, whereas the muscles of the trunk and legs are not affected; in fact, the growth and development of limbs are stimulated as a part of metamorphosis. The effect of the hormone depends on the nature of the reacting cells and tissues—*i.e.*, on their competence—just as the embryonic inductor in the earlier stages of development influences only cells with the competence for a particular kind of reaction.

DIRECT DEVELOPMENT

If an animal after birth or emergence from an egg differs from the adult in comparatively minor details (apart from not having functional sex organs), the development is said to be direct. There is no larval stage and no metamorphosis. Direct development does not mean, however, that no changes occur between birth and adulthood. One very obvious change is the growth of the animal.

The rate of growth—not absolute increase—is highest in the early stages of postembryonic life; subsequently, growth continues to slow, ceasing completely at the attainment of adulthood. The rate of growth is dependent on many factors, both external (feeding, temperature) and internal. Of the internal factors, the most important are hormones, especially the growth hormone produced by the hypophysis. If the growth hormone is produced in insufficient quantities, the result is dwarfism; if it is produced in excessive quantities, the result is gigantism.

Factors influencing rate of growth

In the case of direct development, the most important change is the attainment of sexual maturity, which is achieved in several steps and involves the action of several hormones. The gonad rudiments and rudiments of the supporting parts of the reproductive system remain

inactive long after birth. At the approach of adulthood, however, two sets of hormones come into action: hypophyseal hormones stimulate the gonads to activity, and gonadal hormones (produced by the gonads) cause the supporting sex organs and other sex characters to become fully developed. To become functional, the gonads must be acted upon by secretions from the hypophysis. In immature females the follicle-stimulating hormone, which alone causes the egg follicles and the oocytes to grow, and the luteinizing hormone stimulate the follicle cells to produce the female sex hormone, estrogen, which effects the development of the uterus, the milk glands, and other characteristics of the female sex. In the male, the same hypophyseal hormones are produced, with the result that the testes start to produce sperm and to secrete the male sex hormone, androgen. It appears that the luteinizing hormone is the more active in the male sex, being able to cause both spermatogenesis and androgen secretion. Androgen, in turn, brings about the development of the penis, the descent of the testicles before birth, the appearance of typical male hair growth, and other secondary sex characteristics.

Maturity and death

Sexual maturity and the ensuing reproductive activity mark the pinnacle of development and morphogenesis and, for many animals, herald the end of life. The biological goal of the entire process is achieved with the launching of the next generation, and the life cycle that runs from the formation of gametes by one generation to the formation of gametes by the next generation is completed. In many animals the females die after laying their eggs; the males may have died earlier, after pairing. Indeed, some males (spiders, praying mantises) are eaten by the females immediately after copulation.

The developmental period can only truly be said to end with the termination of an organism, for much activity continues to unfold new developmental sequences, not all of them progressive and favourable, to be sure. Senescence, or a decline in abilities, signals advancing age in mammals but is not a general occurrence in the animal kingdom. Far more animals continue to function at near-peak capacity well into old age. And even among those species—salmons, eels, many moths—whose members die after a single reproductive act, death is relatively swift and not accompanied by a prolonged period of deterioration.

In most animals the reproductive potential is not exhausted in a single act of gamete production, but the sexually mature individuals remain alive and reproduce repeatedly. In these cases life may extend long beyond the first attainment of reproductive ability and be accompanied by further growth of the individuals, as occurs in most fishes, amphibians, and reptiles, and also in mollusks and certain other invertebrates. In the case of prolonged life-spans, however, reproductive activity may cease with advancing age, and a senile involution take place, as is observed mainly in mammals and, particularly, in man. The changes taking place may be described as regressive development. In most animals, however, the end of life is not preceded by any overt traces of senility. As a general rule, then, the attainment of reproductive ability may be said to be the final phase of progressive development among animals.

A gradual loss of alertness and vigour is typical of the aging pattern of primates and is especially important to man. (For additional information, see below *Aging and senescence.*) (B.I.B.)

HUMAN GROWTH AND DEVELOPMENT

The human body, like that of most animals, develops from a single cell produced by the union of a male sex cell and a female sex cell. Human development follows closely the basic vertebrate pattern, and it departs only in certain details from the type specifically characteristic of mammals (see above *Animal development*). A prenatal period, in which most of the developmental advances occur, is followed by a long postnatal period. Only at about the age of 25 are the last progressive changes completed.

Human embryology: early stages

FROM FERTILIZATION TO PLACENTATION

Penetration of the ovum by the sperm and their union

Fertilization. The development and liberation of the male and female gametes are steps preparatory to their union through the process of fertilization. Random movements first bring some spermatozoa into contact with follicle cells adhering to the secondary oocyte, which still lies high in the uterine tube. The sperm then propel themselves past the follicle cells and attach to the surface of the gelatinous zona pellucida enclosing the oocyte. Some sperm heads successfully penetrate this capsule by means of an enzyme, hyaluronidase, which they secrete, but only one makes contact with the cell membrane and cytoplasm of the oocyte and proceeds farther. This is because the invading sperm head releases a substance that initiates surface changes in the oocyte cytoplasm that other competitors cannot master.

The sole successful sperm is engulfed by a conical protrusion of the oocyte cytoplasm and is drawn inward. Once within the periphery of the oocyte, the sperm advances toward the centre of the cytoplasm; the head swells and converts into a typical nucleus, now called the male pronucleus, and the tail detaches. It is during the progress of these events that the oocyte initiates its final maturation division. Following the separation of the second polar body, the oocyte nucleus reconstitutes typically and is then called the female pronucleus of the ripe egg. It is now ready to unite with its male counterpart and thereby consummate the total events of fertilization.

The two pronuclei next approach, meet midway in the egg cytoplasm, and lose their nuclear membranes. Each resolves its diffuse chromatin material into a complete, single set of 23 chromosomes. Centrioles (complex particles involved in cell division), apparently supplied by the sperm, appear, and a mitotic spindle organizes with the two sets of chromosomes arranged midway on it—ready to proceed with a typical mitosis. This climax in the events of fertilization creates a joint product named the zygote. It contains all the essential factors for the development of a new individual.

The fundamental results of fertilization are the following: (1) reassociation of a male and female set of chromosomes (thus restoring the full number and providing the basis for biparental inheritance and for variation); (2) establishment of the mechanism of sex determination for the new individual (this depending on whether the male set of chromosomes included the X- or the Y- chromosome); (3) activation of the zygote, initiating a beginning toward the production of a new individual.

Cleavage and blastulation. *Cleavage.* The onset of mitosis (ordinary cell proliferation by division) in the activated zygote is a first step toward development in the ordinary sense of that term, and the cells so produced are the first external sign of future body building. To this end, the relatively enormous zygote directly subdivides into many smaller cells of conventional size, suitable as early building units for the future organism. This process is called cleavage and the resulting cells are blastomeres (Figures 29A–D). The tendency for the progressive increase in cell numbers to follow a doubling sequence is soon disturbed and then lost. Each blastomere receives the full complement of paternal and maternal chromosomes.

Formation of blastomeres

Subdivision of the zygote into blastomeres begins while it is still high in the uterine tube. The cohering blastomeres are transported downward chiefly, at least, by muscular contractions of the tubal wall. Such transport is relatively rapid until the lower end of the tube is reached, and here cleavage continues for about two days before the multicellular cluster is expelled into the uterus. The full reason for

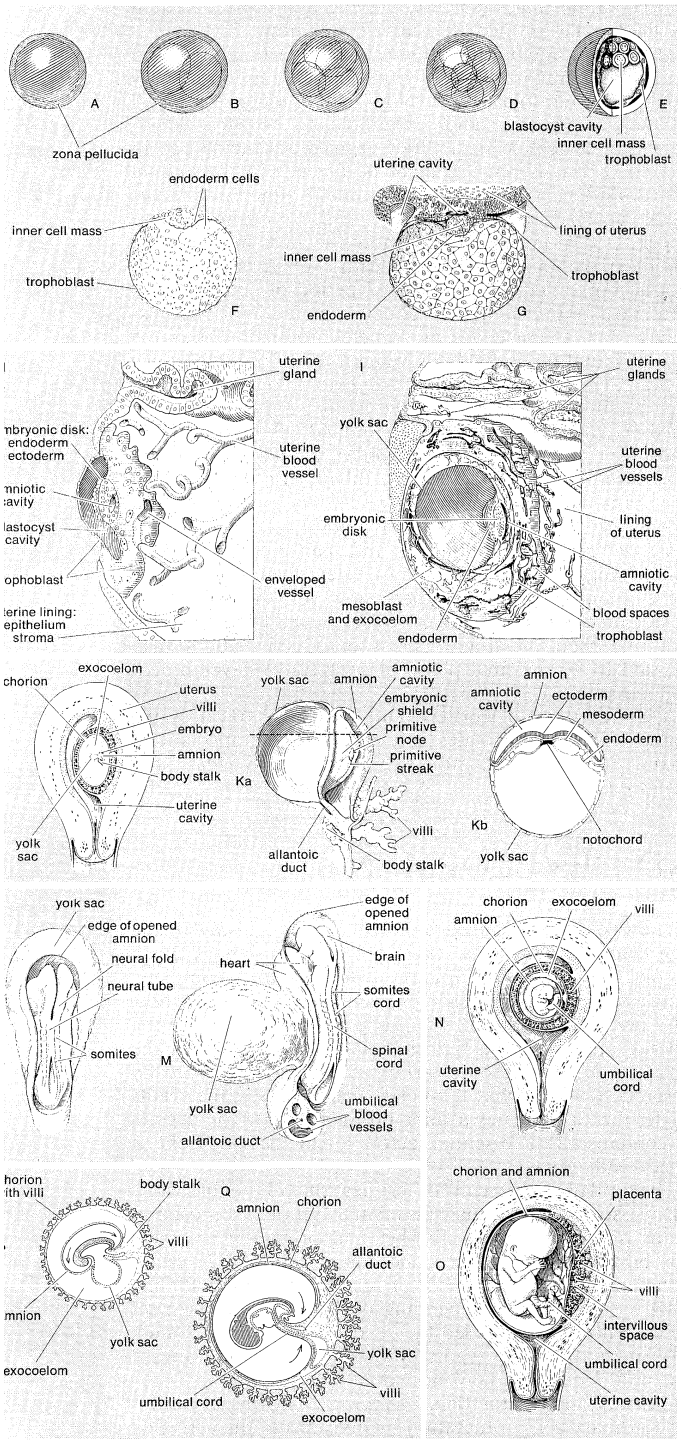


Figure 29: *Development of the human embryo.* A–E (rabbit) and F–G (monkey) illustrate comparable first stages of human development. (A–D) Cleavage of ovum (magnified 80×). (E) Early free blastocyst, partially opened (magnified 80×). (F) Older free blastocyst, halved (magnified 60×). (G) Late attached blastocyst, halved (magnified 70×). (H–J) Half sections of embryos implanted in uterine lining: (H) at 7½ days (magnified 80×); (I) at 13 days (magnified 37×); (J) at 23 days (magnified ¼×). (Ka) Embryo of 18 days at disk (or shield) stage, three-quarter view (magnified 9×). (Kb) Cross section through same embryo at level of broken line in Ka. (L) Embryo of 21 days with amnion opened, back view (magnified 8×). (M) Embryo of 23 days with yolk sac and opened amnion (magnified 9×). (N,O) Embryo of six weeks and fetus of three months, respectively, within halved amnion, chorion, and uterus (magnified ¼×). (P,Q) Diagrams showing especially amnion growth and formation of umbilical cord; stages correspond to J and N.

Drawing by James F. Didusch

this delay is not clear, but it serves to retain the cleaving blastomeres until the uterine lining is suitably prepared to receive its prospective guest.

Since the human egg contains little inert yolk material, and since this is distributed rather evenly throughout the cytoplasm, the daughter cells of each mitosis are practically equal in size and composition. This type of cleavage is known as total, equal cleavage. The sticky blastomeres adhere and the cluster is still retained for a time within the gelatinous capsule—the zona pellucida—that had enclosed the growing and ovulated oocyte. There is no growth in the rapidly dividing blastomeres, so that the total mass of living substance does not increase during the cleavage period. Cleavage is not a mechanism designed primarily to distribute particular developmental qualities, either nuclear or cytoplasmic, to specific blastomeres, which then carry out assignments that give rise to particular parts of the embryo. In a sense, therefore, cleavage is essentially a nonspecific fractionating preliminary rather than a process of truly constructive development.

Passage of blastomeres into uterus; formation of morula and blastocyst

Morula and blastocyst. By the fourth day after ovulation, a cluster of about 12 blastomeres passes from the uterine tube into the uterus. When the cluster numbers 12 to 16 blastomeres it is called a morula (Figure 29D). By the time some 30 blastomeres have been produced, pools of clear fluid accumulate between some of the internal cells, and these spaces soon coalesce into a common subcentral cavity. The resulting hollow cellular ball is a blastula of a particular type that occurs in mammals and is called a blastocyst; its cavity is the blastocoel (Figure 29E, F).

An internal cellular cluster, eccentric in position and now named the inner cell mass, will develop into the embryo. The external capsule of smaller cells, enveloping the segregated internal cluster, constitutes the trophoblast. It will contribute importantly to the formation of a placenta. During its stay within the uterine cavity, the blastocyst loses its gelatinous capsule, imbibes fluid, and expands to a diameter of 0.2 millimetre (0.008 inch); this is nearly twice the diameter of the zygote at the start of cleavage. Probably several hundred blastomeres have formed before the blastocyst attaches to the uterine lining.

Implantation and placentation. Late in the sixth day after ovulation the naked, sticky blastocyst comes into contact with the uterine lining and adheres to it (Figure 29G). The site of attachment is variable and not predetermined. The uterine lining has already been preparing, under the influence of ovarian hormones, for the reception of the blastocyst. Among these preparations has been the elaboration and expulsion, by the uterine glands, of a secretion that serves as nourishment for the blastocyst, both when it is free and during its implantation. Directly after blastocyst attachment come its establishment within the thickened uterine lining and the participation of its trophoblastic capsule in the differentiation of a placenta, a structure that enables the developing embryo to enter into a prompt, physiological dependence on the mother.

Implantation. The trophoblast of the blastocyst exerts an enzymic, destructive influence on the swollen uterine lining, leading to erosion of both the superficial epithelium of the uterine lining and also its deeper, cellular connective tissue (Figure 29H). This early stage of invasion ends in a few days; the blastocyst is then completely buried within a more superficial and compact layer of the total uterine lining (Figure 29H, I). While the blastocyst is completing this phase of implantation, its original shell of cellular trophoblast steadily proliferates a multitude of cells that lose their outermost membranes and merge. The result is a thick peripheral layer consisting of a common mass of cytoplasm in which many nuclei are embedded. This external investment is called syncytial trophoblast.

Syncytial trophoblast

The implanted blastocyst next proceeds to establish itself as a parasite within the uterus. The syncytial trophoblast becomes a spongy shell containing irregular cavities (Figure 29I). This expanding mass destroys cellular connective tissue and glands encountered in its path. Both the cellular and derivative syncytial trophoblast have the capacity of ingesting such tissue. The erosive process also taps capillaries and small blood vessels, and blood liberated in this way is likewise taken up into the trophoblast. Such en-

gulfed blood cells and fragments of tissue are presumably utilized for nourishment in the early period of establishment of the expanding blastocystic sac. Yet these erosive activities decline in intensity by the end of the third week of development, and at this time the sac is completing the first phase of its specialization (Figure 29L).

Ectopic pregnancy

Occasionally a fertilized egg fails to reach the uterus, implanting and beginning to develop elsewhere. This outcome is called an ectopic or extra-uterine pregnancy. The commonest ectopic site is the uterine tube, but the peritoneum lining the abdominal cavity and even the interior of the ovary are also involved, though rarely. The unsuitability of all these sites for continued development usually leads to early death and resorption of the embryo.

Placentation. The irregular strands of invasive syncytial trophoblast constitute a first stage in the formation of true villi (which form part of the placenta and are briefly described below). Primitive connective tissue soon lines the interior of the blastocyst wall, and this complex (trophoblast and connective tissue) is then named the chorion. Connective tissue promptly grows into the trophoblastic strands, and blood vessels develop in the tissue. The result is the production of many chorionic villi, each resembling a tiny, branching bush (Figure 29J, Ka).

In the fourth week of development, the essential arrangements have been established that make possible those physiological exchanges between mother and fetus that characterize the remainder of pregnancy. The deepest embedded portion of the chorionic wall becomes the so-called chorionic plate of the developing placenta. From the plate extend the main stems of chorionic villi, which give off progressively subdividing branches (Figure 29N, O). In general, the side branches are free, whereas apical ones tend to attach to the maternal tissue and serve as anchors. The villous trees occupy a labyrinthine space between the villi that was created by erosion of the uterine lining. Trophoblast not only covers the chorionic plate and its villi but also spreads like a carpet over the eroded surface of the maternal tissue.

Tapped uterine arteries open into the trophoblast-lined intervillous space, their blood bathing the branches and twigs of the villous trees. This blood drains from the intervillous space through similarly tapped veins. Arterial blood of the embryo, and later fetus, passes through vessels of the umbilical cord to the chorionic plate. Thence it is distributed to the villous stems, branches, and twigs through vessels in their connective-tissue cores. Return of this blood to the fetus is by a reverse route.

The circulation of maternal blood through the intervillous space is wholly separate from fetal blood coursing through the chorion and its villi. Communication between the two is solely by diffusive interchange. The barrier between the two circulations consists of the trophoblastic covering of villi, the connective tissue of the villous cores, and the thin lining of the capillaries that are contained in the villous cores.

Functions of the placenta

The placenta serves the fetus in several ways, most of which involve interchanges of materials carried in the bloodstreams of the mother and fetus. These functions are of the following kinds: (1) nutrition, (2) respiration, (3) excretion, (4) barrier action (*e.g.,* prevention of intrusions by bacteria), and (5) synthesis of hormones and enzymes. (For functional details and for a description of the mature, expelled placenta, see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction from conception to birth: The normal events of pregnancy.*)

Extra-embryonic membranes. The way in which the encapsulating membrane of the blastocyst becomes the chorion, and the most deeply embedded part of it becomes the fetal placenta, has already been described. There are still other important membranes that develop from those portions of the inner cell mass of the blastocyst that are not directly involved in becoming an embryo.

Yolk sac. Cells split off from the inner cell mass of the blastocyst and fashion themselves into a primitive yolk sac (Figure 29J, Ka). The roof of the sac then folds into a tubular gut, whereas the remainder becomes a vascularized bag that attains the size of a small pea (Figure 29M). In other vertebrates, such as amphibians and birds, the yolk sac is large and contains a store of nutritive yolk. But in man and other true mammals there is practically none. A slender neck, the yolk stalk, soon connects the rapidly elongating gut with the fast growing yolk sac proper. The stalk detaches from the intestine early in the second month, but the shrunken sac commonly persists and can be found in the expelled afterbirth.

Amnion. A cleft separates the outermost cells of the inner cell mass of the blastocyst from the remainder, which then becomes the embryonic disk (Figure 29H). The split-off, thin upper layer is the amnion, which remains attached to the periphery of the embryonic disk. As the disk folds into a cylindrical embryo, the amniotic margin follows the underfolding, and its line of union becomes limited to the ventral (frontward) body wall, where the umbilical cord attaches (Figure 29P, Q). The amnion becomes a tough, transparent, nonvascular membrane that gradually fills the chorionic sac and then fuses with it (Figure 29N, O). At the end of the third month of pregnancy, the nonplacental extent of this nearly exposed double membrane comes into contact with the lining of the uterus elsewhere. Fusion then obliterates the uterine cavity, which has been undergoing progressive reduction in size. For the remainder of pregnancy the only cavity within the uterus is that of the fluid-filled amniotic sac.

The amniotic sac

Clear, watery fluid fills the amniotic sac. The embryo is suspended in this fluid and thus can maintain its shape and mold its body form without hindrance. Throughout pregnancy the amniotic sac serves as a water cushion, absorbing jolts, equalizing pressures, and permitting the fetus to change posture. At childbirth it acts as a fluid wedge that helps dilate the neck of the uterus. When the sac ruptures, about a quart of fluid escapes as the "waters." If the sac does not rupture or if it covers the head at birth, it is known as a caul.

Allantois. The allantois, a tube of endoderm (the innermost germ layer), grows out of the early yolk sac in a region that soon becomes the hindgut. The tube extends into a bridge of mesoderm (the middle germ layer) that connects embryo with chorion and will become incorporated into the umbilical cord (Figure 29P, Q). The human allantoic tube is tiny and never becomes a large sac with important functions, as it does in reptiles, birds, and many other mammals. In the second month it ceases to grow, and it soon is obliterated. Blood vessels, however, develop early in its mesodermal sheath, and these spread into the chorion and vascularize it. Throughout pregnancy they will keep the embryo in close relationship with the mother's uterine circulation.

Umbilical cord. As the ventral body wall closes in, the yolk stalk and allantois are brought together, along with their mesodermal sheaths and blood vessels (Figure 29P, Q). Enclosing everything is a wrapping of amnion. In this manner a cylindrical structure, the umbilical cord, comes to connect the embryo with the placenta (Figure 29N, O). It will serve the embryo and fetus as a physiological lifeline throughout the period of pregnancy. The mature cord is about 1.3 centimetres (0.5 inch) in diameter, and it attains an average length of nearly 50 centimetres (two feet).

FORMATION OF THE THREE PRIMARY GERM LAYERS

The inner cell mass, attached to the deep pole of the implanted blastocyst, is sometimes called the embryoblast, since it supplies the materials used in the formation of an embryo. The cellular mass flattens and enters into the process of gastrulation, through which the three primary germ layers segregate and the gastrula stage, the next advance after the blastula, begins to take form.

Gastrulation

First, cells facing the cavity of the blastocyst arrange into a layer named the endoderm (Figure 29H, I). The thick residual layer, temporarily designated as epiblast, is the source of a definitive uppermost sheet, the ectoderm, and an intermediate layer, the mesoderm. In this second phase of gastrulation, some cells of the epiblast migrate to the midline position, then turn downward and emerge beneath as mesoderm. Such cells continue to spread laterally, right and left, between the endoderm and the residue of epiblast, which is now definitive ectoderm (Figure 29Kb).

The site where the migratory mesodermal cells leave the epiblast is an elongated, crowded seam known as the

primitive streak (Figure 29Ka). Similar migrating cells produce a thick knob at one end of the primitive streak. Their continued forward movement from this so-called primitive knot produces a dense band that becomes the rodlike notochord.

The germ layers are not segregated sheets whose cells have predetermined, limited capacities and inflexibly fixed fates in carrying out organ-building activities. Rather, the layers represent advantageously located assembly grounds out of which the component parts of the embryo emerge normally, according to a master constructional plan that assigns different parts to definite spatial positions and local sites. Thus, although the germ layers have developmental potencies in excess of their normal developmental fates, their ordinary participation in organ forming does not deviate from a definite, standard program.

The derivatives of the primary germ layers can best be presented in tabular form. In naming the germ-layer origin of an organ, only the principal functional tissue is designated (Table 1). In a few instances, such as the suprarenal (adrenal) glands and the teeth, a compound organ has important parts of different origin.

Table 1: Derivatives of Primary Germ Layers

ectoderm	mesoderm	endoderm
epidermis	epithelium of:	epithelium of:
cutaneous derivatives	circulatory system	pharynx
epithelium of:	spleen; lymph nodes	thyroid; thymus
mouth; oral glands	urogenital system	parathyroid
nasal passages	body cavities	digestive tube;
sense passages	connective tissues;	liver; pancreas
central nervous system	blood; bone marrow	larynx; trachea; lungs
peripheral nervous system	muscular tissues	urinary bladder; urethra
hypophysis;	skeletal tissues	vestibule; vagina
suprarenal medulla	suprarenal cortex	

GROWTH AND DIFFERENTIATION

Growth. Growth is an increase in size, or bulk. Cell multiplication is fundamental to an increase in bulk but does not, by itself, result in growth. It merely produces more units to participate in subsequent growing. Growth is accomplished in several ways. Most important is synthesis, by which new living matter, protoplasm, is created from available foodstuffs. Another method utilizes water uptake; a human embryo of the early weeks is nearly 98 percent water, while an adult is 70 percent fluid. A third method of growth is by intercellular deposition; cells manufacture and extrude such nonliving substances as jelly, fibres, and the ground substance of cartilage and bone. Through these activities a newborn baby is several thousand million times heavier than the zygote from which it came.

Ways in which growth is accomplished

It is obvious that uniform growth throughout the substance of a developing organism would merely produce a steadily enlarging spherical cellular mass. Local diversities in form and proportions result from differential rates of growth that operate in different regions and at different times. The particular program of starting times and growth rates, both externally and internally in the human embryo, constitutes its characteristic growth pattern. Abnormal growth occurs occasionally, and growth may be excessive or deficient. Also, such departures may be general or local, symmetrical or asymmetrical. General gigantism usually starts before birth, and the oversized baby continues to grow at an accelerated rate. On the other hand, an existing hereditary predisposition may not be aroused into action until some time during childhood. In a reverse manner, general dwarfism may exist before birth, the individual continuing to grow slowly and stopping at the usual time. Another type is normal in size at birth, grows normally for a while, and then comes to a premature arrest.

Differentiation. In a developing organism, differentiation implies increasing structural and chemical complexity. One kind of differentiation concerns changes in gross shape and organization. Such activities, related to molding the body and its integral parts into form and pattern, comprise the processes called morphogenesis. The processes of morphogenesis are relatively simple mechanical acts: (1) cell migration; (2) cell aggregation, forming masses, cords, and sheets; (3) localized growth or retardation, resulting, in enlargements or constrictions; (4) fusion; (5) splitting, including separation of single sheets into separate layers, formation of cavities in cell masses, and forking of cords; (6) folding, including circumscribed folds that produce inpocketings and outpocketings; (7) bending, which, like folding, results from unequal growth.

Morphogenesis and histogenesis

A second kind of differentiation refers to progressive changes occurring in the substance and structure of cells, whereby different kinds of tissues are created. These changes, and the synthetic processes underlying them, constitute histogenesis. The zygote contains all the essential factors for development, but they exist solely as an encoded set of instructions localized in the genes of chromosomes and bearing no direct physical relationship to the future characteristics of the developing embryo. During histogenesis these instructional blueprints are decoded and transformed, through cytoplasmic syntheses, into the several types and subtypes of tissues that are the structural and functional units of organs. At first, the cells of each germ layer lack an identifiable shape and are similar in chemical composition, but an invisible type of chemical differentiation soon enters. After the elaboration of specific enzyme patterns and syntheses, certain groups of cells progressively assume distinctive characters that permit their fates to be recognized. Such early stages in definite lines of differentiation of cells are often designated by the suffix "-blast" (*e.g.*, myoblast; neuroblast).

The emerging cell types are discrete entities, without intermediates; for example, a transitional form between a muscle cell and a nerve cell is never seen. Neither can different, local parts of a cell carry out different types of tissue specialization, such as nerve at one end and muscle at the other end; nor can a cell, once fully committed to a particular type of specialization, abandon it and adopt a new course.

Under certain conditions, differentiated cells may reverse their course and return to a simpler state. Thus, under a changed environment, cartilage may lose its matrix, and its cells may come to resemble the more primitive tissue from which it arose. Nevertheless, despite such reversal and apparent simplification ("dedifferentiation"), these cells retain their former histological specificity. Under suitable environmental conditions they can differentiate again but can only regain their previous definitive characteristics as cartilage cells.

"Dedifferentiation"; the four types of tissue

The final result of histogenesis is the production of groups of cells similar in structure and function. Each specialized group constitutes a fundamental tissue. There are four main types of such tissues: each of the three germ layers gives rise to sheetlike epithelia, which cover surfaces, line cavities, and are frequently glandular; ectoderm also forms the nervous tissues; mesoderm also produces the muscular tissues; it also differentiates into blood and the fibrous connective tissues (including two further specialized types, cartilage and bone).

EMBRYONIC ACQUISITION OF EXTERNAL FORM

At the end of the second week, the embryonic region is a nearly circular plate within its well-embedded, differentiating chorionic sac (Figure 29I). This embryonic disk consists of two layers (epiblast and endoderm), with a third layer (mesoderm) just starting to spread between them. A hollow, dome-shaped amnion sac attaches to the margin of the upper layer of the disk, and a hollow yolk sac is similarly continuous with the lower layer. A broad cellular bridge attaches the complex to the chorion.

Early in the third week the embryonic disk has enlarged and become pear-shaped in outline, and a well-formed primitive streak occupies the midline of its caudal (hind) half, which is narrower. Cells from the epiblast are passing through the streak and spreading laterally in both directions beneath the uppermost layer, now ectoderm (Figure 29Ka, Kb). In this way the embryonic disk becomes three layered, and the gastrula stage of development comes to an end. At the middle of the week a thickening, the head process, is extending forward from a knoblike primitive knot located at the head end of the primitive streak. These linear thickenings define the median plane of the future

embryo and thus divide the embryonic disk into precise right and left halves.

Toward the end of the week the disk elongates and becomes slipper-shaped in outline; a slight constriction demarcates it from the attached yolk sac. Growth has lengthened the region ahead of the now receding primitive streak. Here, in the midline, the ectoderm bears a definite gutter-like formation called the neural groove; it is the first indication of the future central nervous system. Beneath the groove the mesodermal head process presently rounds into an axial rod, the notochord, that serves as a temporary "backbone." By the end of the third week a head fold, paired lateral body folds, and a tail fold become prominent, demarcating a somewhat cylindrical embryo from the still broadly attached yolk sac (Figure 29L). The neural folds, flanking the neural groove, converge and begin to meet midway of their lengths, thereby producing a neural tube at that level. Mesoderm, alongside the notochord, begins to subdivide into paired blocks called somites, and the outlines of the somites show externally. From them, muscles and vertebrae will differentiate later. This stage, when the embryo is fashioning a neural tube, is often designated as a neurula.

In the fourth week the embryo goes beyond the external characteristics of vertebrates in general and becomes recognizable as a mammal. The week is marked by profound changes, during which the embryo acquires its general body plan. There is an increase in total length from about two to five millimetres, but size is quite variable among smaller specimens. Better correlated with the degree of development is the number of mesodermal somites, which attain their full number of about 42 during the fourth week. Some of the head of an early embryo arises from the embryonic disk in front of the primitive knot. But as the primitive streak shortens and its caudal retreat continues, such structures as the neural tube and notochord are added progressively in the wake of that retreat, and additional somite pairs also appear in steady succession.

Factors in formation of early embryo

The most important manoeuvre in the establishment of general body form is the transformation of the flat embryonic disk into a roughly cylindrical early embryo, which is attached to the yolk sac by a slender yolk stalk (Figure 29M). Three factors cooperate in producing this change: (1) There is more rapid expansion of both the embryonic area and the yolk sac than in the region joining the two. The enlarging embryonic area at first buckles upward

By courtesy of the Carnegie Institution of Washington

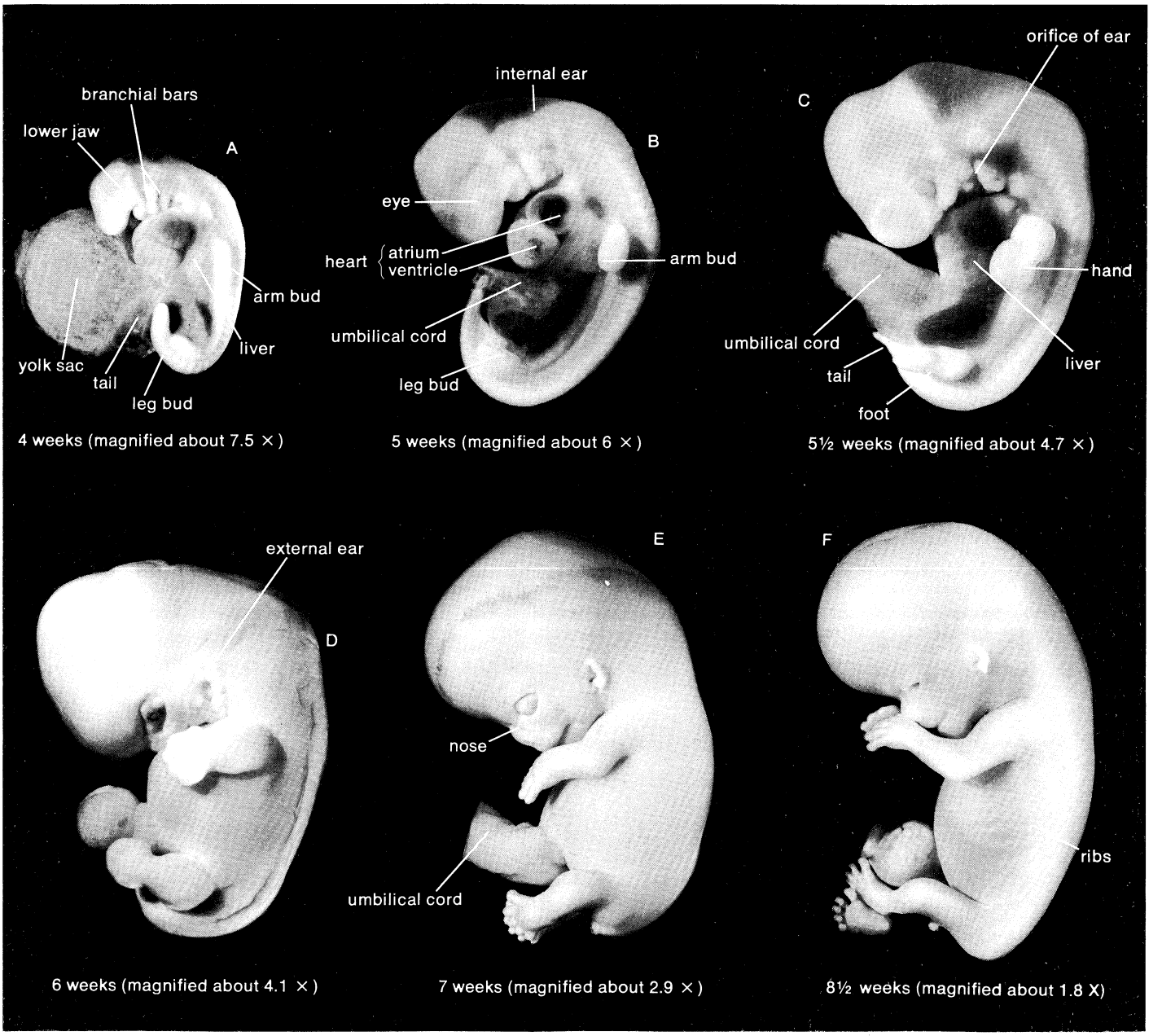


Figure 30: Human embryos from four to 8½ weeks.

and then overlaps the more slowly growing margin. Since growth is particularly rapid at the future head end and tail end, the embryo becomes elongate. (2) In conjunction with this overgrowth there is important underfolding, again most pronounced at the front and hind ends. Underfolding is produced by differential elongation in the regions of the brain and tail bud. Conspicuous is the change in the future cardiac (heart) and foregut region, which swings beneath the brain as on a hinge. (3) A certain amount of true constriction, through growth, purses all of these parts at the site of the future umbilicus.

Throughout the entire period when the body and its parts are being laid down, developmental advances tend to appear first at the head end and then progress tailward. For this reason, many structures that extend along the body for a distance show a gradation in development. The size advantage gained initially by the head end of the embryo is relinquished very slowly. Even in an infant the relatively large head and long arms are striking. A further tendency toward progressively graded development occurs from the middorsal line in a lateral (sideward) and then ventral (frontward) direction. All such relations are the visible expressions of stages in growth and differentiation.

Early in the fourth week the cylindrical shape of the embryo is plain, even though the folding-off process is far from complete (Figure 29M). The neural tube is still open near both ends, and at the head end the broader neural folds indicate the future brain and even its three primary divisions. A pronounced bulge beneath the brain region denotes where the heart is forming precociously in order to institute a necessary, prompt circulation through the placenta.

During the middle and late days of the fourth week there are marked advances (Figure 30A). Accelerated growth along the dorsal region bends the total body length progressively until the embryo assumes a striking C-shape, with the tips of the head and tail not far apart. Continued growth and underfolding close in much of the ventral side of the embryo, so that a free head and upper trunk, and a lower trunk and a prominent tail, are easily recognizable. Forebrain, midbrain, and hindbrain can be identified, largely because of a sharp bend in the midbrain. Local outgrowths from each side of the forebrain produce stalked eye cups, and a pair of inpocketings of the ectoderm alongside the hindbrain sink beneath the surface as otic vesicles, forerunners of the inner ears. Bulges indicative of the heart and liver are prominent. Formations called branchial arches, reminiscent of the gill arches of fishes and aquatic amphibians, become conspicuous in the future jaw–neck region. Paired swellings ("buds") off the trunk foretell the locations of the upper and lower limbs.

Changes in the fifth to eighth weeks

A five-week embryo is about eight millimetres (0.32 inch) long (Figure 30B, C), whereas at six weeks the length is about 13 millimetres (0.52 inch) (Figure 30D). New external features are olfactory pits at the tips of the bent head. An umbilical cord becomes a definite entity, its proximal end occupying a low position on the abdominal wall. The sharply bent head joins the rest of the body at an acute angle. The first pair of branchial arches branch Y-fashion into maxillary and mandibular processes (primitive upper and lower jaws). The external ears are forming around the paired grooves located between each half of the mandible and each second branchial arch. The heart, which was previously the chief ventral prominence, now shares this distinction with the rapidly growing liver. Limb buds have elongated markedly and become flattened at their outer ends. A constriction on each bud separates a paddle-like hand plate or foot plate from a cylindrical segment attached to the body wall. Predictably, the upper limbs are somewhat further advanced than the lower pair.

In the latter weeks of the second month, developmental changes advance from those that distinguish primates to a state that is recognizably human (Figure 30E). At the end of the second month (30 millimetres', or 1.2 inches', length) the stage of the embryo ends; henceforth, until birth, "fetus" is the preferred term.

The fetal stage

In the seventh and eighth weeks the head becomes more erect and the previously curved trunk becomes straighter (Figure 30E, F). The heart and liver, which earlier dominated the shape of the ventral body, yield to a more evenly rounded chest–abdomen region. The tail, which at an earlier time was one fifth of the embryo's length, becomes inconspicuous both through actual regression and through concealment by the growing buttocks. The face rapidly acquires a fairly human appearance; eyes, ears, and jaws are prominent. The eyes, previously located on the sides of the head, become directed forward; the nose lacks a bridge and so is of the "pug" type, with the nostrils directed forward instead of downward. A mandibular branch of each Y-shaped branchial arch combines with its mate to form the lower jaw. The maxillary branch on each side joins an elevation on the medial (inner) side of the corresponding nostril to produce the more complicated upper jaw. Branchial arches, other than those forming the jaws and external ears, are effaced through incorporation into an emerging, recognizable neck. Limbs become jointed, and the earlier hand plates and foot plates differentiate terminal digits. Primitive external genitalia appear, but in a nondistinctive, sexless condition.

Almost all of the internal organs are well laid down at the end of eight weeks, when the embryo is little more than 25 millimetres (one inch) long. The characteristic external features are established, and subsequent growth merely modifies existing proportions without adding new structure. Similarly, the chief changes undergone by internal organs and parts are those of growth and tissue specialization. At eight weeks the neuromuscular mechanism attains a degree of perfection that permits some response to delicate stimulation.

Development from the third month to birth

During the third month the young fetus clearly resembles a human being, although the head is disproportionately large (Figure 29O). The previous protrusion of much of the intestine into the umbilical cord is reduced through the return of its loops into the abdomen. The ears rise to eye level and the eyelids fuse shut. Nails begin forming; ossification (bone-forming) centres appear in most of the future bones; and the sex of external genitalia becomes recognizable. (In this paragraph, and in the next two, the months are lunar months, of 28 days).

At four months individual differences between the faces of fetuses become distinguishable. The face is broad but the eyes are now less widely separated. The umbilical cord attaches higher on the abdominal wall; this location is above an expanding region between the cord and the pubis (front bones of the pelvis) that scarcely existed previously.

At five months downy hairs (lanugo) cover the body, and some head hairs appear. The skin is less transparent. Fetal movements ("quickening") are felt by the mother. At six months eyebrows and eyelashes are clearly present. The body is lean, but its proportions have improved. The skin is wrinkled. At seven months the fetus resembles a dried-up old person. Its reddish, wrinkled skin is smeared with a greasy substance (vernix caseosa). The eyelids reopen. At eight months fat is depositing beneath the skin. The testes begin to invade the scrotum. At nine months the dull redness of the skin fades and wrinkles smooth out. The body and limbs become better rounded.

At full term (38 weeks) the body is plump and proportions are improved, although the head is large and the lower limbs are still slightly shorter than the upper limbs. The skin has lost its coat of lanugo hair, but it is still smeared with vernix caseosa. Nails project beyond the finger tips and to the tips of toes. The umbilical cord now attaches to the centre of the abdomen. The testes of

Table 2: Age, Size, and Weight

age from date of conception (calendar months)	sitting height		weight	
	mm	in.	g	lb
Two	28	1	2.25	.75 oz
Three	75	3	25	1 oz
Four	135	5.3	170	6 oz
Five	185	7.3	440	14 oz
Six	225	9	820	1.75 lb
Seven	270	10.6	1,380	3 lb
Eight	310	12.2	2,220	5 lb
Nine	360	14.2	3,150	7 lb

males are usually in the scrotum; the greater lips of the female external genitalia, which previously gaped, are now in contact. Cranial bones meet except at some angular junctions, or "soft spots."

The average time of delivery is 280 days from the beginning of the last menstrual period, whereas the duration of pregnancy (age of the baby) is about 266 days (38 weeks; Table 2). Pregnancy may extend to 300 days, or even more, in which case the baby tends to be heavier. Premature babies born under 27 weeks of age rarely survive, whereas those more than 30 weeks old usually do survive. (For additional information on the events of delivery, see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *Human reproduction from conception to birth: Parturition.*)

Abnormal development

MULTIPLE BIRTHS

It is both unusual and abnormal for the human species to produce more than one offspring at a time. "Twins" and "twinning" are used as general terms for any number of multiple births, as the same basic principles apply.

Fraternal and identical twins

Fraternal twins stem from multiple ovulations in the same cycle. Each oocyte develops singly in a separate follicle, is shed and fertilized individually, develops within its own chorionic sac, and forms an individual placenta. In some instances, two blastocysts implant close together and the expanding placentas meet and fuse. In such double placentas, however, the two blood circulations rarely communicate. The word dizygotic technically designates two-egg twins. Such pairs obviously are independent in sex determination and bear no more resemblance than do other children of the same parents. Properly speaking, they are merely littermates. Nearly three-fourths of all American twins are dizygotic, whereas the Japanese ratio is only one-fourth. A tendency toward such multiple births exists in some family lines.

Wholly different are those true twins who are always of the same sex and are strikingly similar in physical, functional, and mental traits. Such close identity is enforced by their derivation from a single ovulated and fertilized egg, and hence by their acquisition of identical chromosomal constitutions. This twin type is named monozygotic. Three-fourths of such pairs develop within a common chorionic sac and share a placenta; one-fourth have individual sacs and placentas. The latter condition results from a mishap before implantation, when the cleavage cells separate into two groups and then become individually implanting blastocysts. There is no discernible hereditary tendency toward the production of monozygotic twins.

Several atypical processes involving the inner cell mass or embryonic plate can produce separate embryos within a single sac: (1) The inner cells of a blastocyst may segregate into two masses. (2) Somewhat later in time, two embryonic axes may become established on a single embryonic disk. (3) A single axis may subdivide by fission or budding. (4) Duplication of any sort may combine with secondary subdivision; the Dionne quintuplets are believed to have followed this sequence, which is also normal for the regular quadruplets of the Texas armadillo.

Conjoined twins

The abnormal nature of monozygotic twinning is emphasized when the two individuals are conjoined as a "double monster." The degree of union varies from slight to extensive, and the possession of a single or double set of internal organs depends on the intimacy of fusion at any particular level. Union occurs by the heads, upper trunks, or lower trunks; the joining may be by the dorsal, lateral, or ventral surfaces. Sometimes there is a marked disparity in the size of the two components, and in such instances the smaller twin is called a parasite. Conjoining results from divergent growth at the front or hind end of the emerging primitive axis of an embryo, or at both ends.

FETAL DEVIATIONS

Human embryos are subject to disease, abnormal development, and abnormal growth. Decline and death can occur at any stage, but since most deaths occur in the first two or three weeks of development they usually escape notice. Probably little more than half of all zygotes reach full-term birth. Most abnormalities resulting from faulty development originate within the first seven weeks of pregnancy, before the prospective mother is surely aware of her condition. Abnormalities that do occur in living infants tend toward the milder types, since the severe mishaps commonly terminate development before birth.

Folklore maintains that a pregnant woman may "mark" her baby through incurring physical injuries or becoming subjected to horrors or repulsive sights. As there are no nervous connections between mother and fetus, such beliefs lack foundation. Moreover, practically all of the alleged causative experiences occur long after the "related" abnormalities have been established in the embryo.

Defective health of the mother can, in some instances, become a cause of the physical impairment or death of a fetus. Certain infectious diseases, for example, may result in fetal injury; such related causative organisms can be a virus (German measles), a spirochetal microorganism (syphilis), or a protozoon parasite (toxoplasmosis). Also, placental disorders, malformations of the mother's reproductive organs, and inadequate functioning of her endocrines may provide an unfavourable environment for normal development. Birth itself imposes the risk of oxygen deficiency or other injury; either may result in some malfunctioning of the brain.

TERATOLOGY

Teratology is concerned with all features of abnormal generation and development of the embryo (embryogenesis) and their end products. The incidence of defective development is high. One infant in 14 that survive the neonatal period bears an abnormality of some kind and degree, and half of these babies have more than one malformation. Internal, concealed defects are more numerous than external ones, and some defects do not become apparent until childhood. One baby in 40 is born with a structural defect that needs treatment. Some types of abnormality are commoner in males (*e.g.,* pyloric stenosis, the narrowing of the opening between the stomach and the intestine), while other types predominate in females (*e.g.,* dislocated hip). Besides the obvious derangements of a gross nature, there are aberrations at the molecular level known as inborn errors of metabolism. In these an enzyme deficiency blocks the course of intermediary metabolism and results in abnormal chemical functioning. Such errors involve proteins, carbohydrates, lipids, and pigments. The abnormal products may be stored or excreted.

Incidence, causation, and types of abnormality

Important among causes of abnormalities are hereditary factors. Such include gene mutations, which may become Mendelian dominants (*e.g.,* fused fingers, which need be inherited from only one parent to appear in the offspring), recessives (*e.g.,* albinism, which does not become evident unless its gene is inherited from both parents), or sex-linked factors (*e.g.,* hemophilia). Besides the heritable defects, whose possibilities of recurrence can be estimated, there are many genetic results that are due to chance, are not passed on, and do not occur in other offspring. An unequal distribution of chromosomes during meiosis, leading to abnormal assortments, occurs in somatic (non-sex) chromosomes (*e.g.,* Down's syndrome) and in sex chromosomes (*e.g.,* Klinefelter's syndrome).

Environmental factors, both external and internal, are also important. Among physical agents, mechanical pressures or blows are no longer considered to be significant because of the protection supplied by the uterus and the fluid-filled amniotic sac. On the other hand, irradiation is a wholly effective physical agency, as experiments have amply proved. Various chemical agents, used experimentally on pregnant animals, are highly teratogenetic (productive of physical defects within the uterus), and at least one drug (thalidomide) has demonstrated its potency on human embryos. Deficiencies of some fetal hormones are associated causally with gross bodily defects (*e.g.,* male hormone and false hermaphroditism, a condition in which the gonads are of one sex but some appearances suggest the other). Similarly, hormonal excess can be productive (*e.g.,* growth-promoting hormone and gigantism). Vitamin deficiencies are effective experimentally but are not factors in ordinary substandard maternal diets. The role of infectious agents has been men-

tioned previously. Immunity incompatibilities may exist between mother and fetus (*e.g.*, Rh blood factor and fetal hemolysis, the release of hemoglobin from red blood cells).

There appear to be several ways in which teratogenetic agents can affect susceptible embryonic cells. But whatever the manner of interference may be, the final result is probably either cell impairment or death or a changed rate of growth. Either of these measures puts local development out of step with adjoining parts and upsets the coordinated schedule of development.

Development of organs

ECTODERMAL DERIVATIVES

Origin of the skin

Integumentary system. The skin has a double origin. Its superficial layer, or epidermis, develops from ectoderm. The initial single-layered sheet of epithelial cells becomes multilayered by proliferation, and cells nearer the surface differentiate a horny substance. Pigment granules appear in the basal layer, at least, of all races. The epidermis of the palm and sole becomes thicker and more specialized than elsewhere. Cast-off superficial cells and downy hairs mingle with a greasy glandular secretion and smear the skin in the late fetal months; the pasty mass is called vernix caseosa. The deep layer of the skin, or dermis, is a fibrous anchoring bed differentiated from mesoderm. In the later fetal months the plane of union between epidermis and dermis becomes wavy. The permanently ridged patterns are notable at the surface of the palm and sole.

Nails, hairs, and skin glands

Nails develop in pocket-like folds of the skin near the tips of digits. During the fifth month specialized horny material differentiates in proliferating ectodermal cells. The resulting nail plate is pushed forward as new plate substance is added in the fold. Fingernails reach the finger tips one month before birth. Hairs, produced only by mammals, begin forming in the third month as cylindrical pegs that grow downward from the epidermis into the dermis. Cells at the base of the peg proliferate and produce a horny, pigmented thread that moves progressively upward in the axis of the original cylinder. This first crop of hairs is a downy coat named lanugo. It is prominent by the fifth month but is mostly cast off before birth. Unlike nails, hairs are shed and replaced periodically throughout life.

Sebaceous glands develop into tiny bags, each growing out from the epithelial sheath that surrounds a hair. Their cells proliferate, disintegrate, and release an oily secretion. Sweat glands at first resemble hair pegs, but the deep end of each soon coils. In the seventh month an axial cavity appears and later is continued through the epidermis. The mammary glands, peculiar to mammals, are specialized sweat glands. In the sixth week a thickened band of ectoderm extends between the bases of the upper and lower limb buds. In the pectoral (chest) region only, gland buds grow rootlike into the primitive connective tissue beneath. During the fifth month 15 to 20 solid cords foretell the future ducts of each gland. These later canalize and open into a pit that after birth elevates into a nipple. Until late childhood the mammary glands are identical in both sexes.

Mouth and anus. *Mouth.* The mouth is a derivative of the stomodaeum, an external pit bounded by the overjutting primitive nasal region and the early upper and lower jaw projections. Its floor is a thin membrane, where ectoderm and endoderm fuse. Midway in the fourth week this membrane ruptures, making continuous the primitive ectodermal mouth and endodermal pharynx (throat). Lips and cheeks arise when ectodermal bands grow into the mesoderm and then split into two sheets. Teeth have a compound origin: the cap of enamel develops from ectoderm, whereas the main mass of the tooth, the dentin, and the encrusting cementum about the root differentiate from mesoderm. The salivary glands arise as ectodermal buds that branch, bushlike, into the deeper mesoderm. Berrylike endings become the secretory acini (small sacs), while the rest of the canalized system serves as ducts. The palate is described in relation to the nasal passages. A tiny pocket detaches from the ectodermal roof of the stomodaeum and becomes the anterior, or frontward, lobe of the hypophysis, also called the pituitary. The anterior lobe fuses with the neural lobe of the gland (see below).

Anus. A double-layered, oval membrane separates the endodermal hindgut from an ectodermal pit, called the proctodaeum, the site of the future anal canal and its orifice, the anus. Rupture at eight weeks creates a communication between the definitive anus and the rectum.

Origin, histogenesis, and morphogenesis of central nervous system

Central nervous system. Both the brain and the spinal cord arise from an elongate thickening of the ectoderm that occupies the midline region of the embryonic disk. The sides of this neural plate elevate as neural folds, which then bound a gutter-like neural groove (Figure 29L). Further growth causes the folds to meet and fuse, thereby creating a neural tube. The many-layered wall of this tube differentiates into three concentric zones, first indicated in embryos of five weeks. The innermost zone, bordering the central canal, becomes a layer composed of long cells called ependymal cells, which are supportive in function. The middle zone becomes the gray substance, a layer characterized by nerve cells. The outermost zone becomes the white substance, a layer packed with nerve fibres. The neural tube also is demarcated internally by a pair of longitudinal grooves into dorsal and ventral halves. The dorsal half is a region associated with sensory functioning and the ventral half with motor functioning.

Neural tube layers

The gray substance contains primitive stem cells, many of which differentiate into neuroblasts. Each neuroblast becomes a neuron, or a mature nerve cell, with numerous short branching processes, the dendrons, and with a single long process, the axon. The white substance lacks neuroblasts but contains closely packed axons, many with fatty sheaths that produce the whitish appearance. The primitive stem cells of the neural tube also give rise to nonnervous cells called neuroglia cells.

Brain. The head end of the neural plate becomes expansive even as it closes into a tube. This brain region continues to surpass in size the spinal cord region. Three enlargements are prominent, the forebrain, midbrain, and hindbrain. The forebrain gives rise to two secondary expansions, the telencephalon and the diencephalon. The midbrain, which remains single, is called the mesencephalon. The hindbrain produces two secondary expansions called the metencephalon and the myelencephalon.

The telencephalon outpouches, right and left, into paired cerebral hemispheres, which overgrow and conceal much of the remainder of the brain before birth. Late in fetal life the surface of the cerebrum becomes covered with folds separated by deep grooves. The superficial gray cortex is acquired by the migration of immature nerve cells, or neuroblasts, from their primary intermediate position in the neural wall. The diencephalon is preponderantly gray substance, but its roof buds off the pineal body, which is not nervous tissue, and its floor sprouts the stalk and neural (posterior) lobe of the pituitary. The mesencephalon largely retains its early tubular shape. The metencephalon develops dorsally into the imposing cerebellum, with hemispheres that secondarily gain convolutions clothed with a gray cortex. The myelencephalon is transitional into the simpler spinal cord. Roof regions of the telencephalon, diencephalon, and myelencephalon differentiate vascular choroid plexuses (including portions of the pia matter, or innermost brain covering, that project into the ventricles, or cavities, of the brain), which secrete cerebrospinal fluid.

Spinal cord. For a time the spinal cord portion of the neural tube tapers gradually to an ending at the tip of the spine. In the fourth month it thickens at levels where nerve plexuses, or networks, supply the upper and lower limbs; these are called the cervical and lumbosacral enlargements. At this time the spine begins to elongate faster than the spinal cord. As a result, the caudal (hind) end of the anchored cord becomes progressively stretched into a slender, nonnervous strand known as the terminal filament. Midway in the seventh month the functional spinal cord ends at a level corresponding to the midpoint of the kidneys. Both the brain and the spinal cord are covered with a fibrous covering, the dura mater, and a vascular membrane, the pia-arachnoid. These coverings differentiate from local, neighbouring mesoderm.

Peripheral nervous system. In general, each craniospinal nerve has a dorsal (posterior) root that bears a ganglion (mass of nerve tissue) containing sensory nerve cells and

Histogenesis of peripheral nervous system

their fibres, and a ventral (anterior) root that contains motor nerve fibres but no nerve cells. Ganglion cells differentiate from cells of the neural crest, which is at first a cellular band pinched off from the region where each neural fold continues into ordinary ectoderm. Each of these paired bands breaks up into a series of lumps, spaced in agreement with the segmentally arranged mesodermal somites. Neuroblasts within these primordial ganglia develop a single stem and hence are called unipolar. From this common stem one nerve process, or projection, grows back into the adjacent sensory half of the neural tube; another projection grows in the opposite direction, helping to complete the dorsal root of a nerve. Neuroblasts of motor neurons arise in the ventral half of the gray substance of the neural tube. They sprout numerous short, freely branching projections, the dendrons, and one long, little-branching projection, the axon; such a neuron is called multipolar. These motor fibres grow out of the neural tube and constitute a ventral root. As early as the fifth week they are joined by sensory fibres of the dorsal root and continue as a nerve trunk.

Cells of the neural crest differentiate into things other than sensory neurons. Among these variants are cells that encapsulate ganglion cells and others that become neurolemma cells, which follow the peripherally growing nerve fibres and ensheath them. The neurolemma cells cover some nerve fibres with a fatty substance called myelin.

Spinal nerves. Spinal nerves are sensorimotor nerves, with dorsal and ventral roots. A network called a brachial plexus arises in relation to each upper limb and a lumbosacral plexus in relation to each lower limb. The spine, elongating faster than the spinal cord, drags nerve roots downward, since each nerve must continue to emerge between the same two vertebrae. Because of their appearance, the obliquely coursing nerve roots are named the cauda equina, the Latin term for horse's tail.

Cranial nerves. Cranial nerves V, VII, IX, and X arise in relation to embryonic branchial arches but have origins similar to the spinal nerves. The olfactory nerves (cranial nerve I) are unique in that their cell bodies lie in the olfactory epithelium (the surface membrane lining the upper parts of the nasal passages), each sending a nerve fibre back to the brain. The so-called optic nerves (II) are not true nerves but only tracts that connect the retina (a dislocated portion of the brain) with the brain proper. Nerves III, IV, VI, and XII are pure motor nerves that correspond to the ventral roots of spinal nerves. The acoustic nerves (VIII) are pure sensory nerves, each with a ganglion that subdivides for auditory functions and functions having to do with equilibrium and posture; they correspond to dorsal roots. Nerves X and XI are a composite of which XI is a motor component.

Autonomic nervous system. The autonomic nervous system is made up of two divisions, the sympathetic and the parasympathetic nervous systems; it controls such involuntary actions as constriction of blood vessels. Some

Origins of autonomic nervous system

cells of the neural crests migrate and form paired segmental masses alongside the aorta, a principal blood vessel. Part of the cells become efferent multipolar ganglion cells (cells whose fibres carry impulses outward from ganglions, or aggregates of nerve cells) and others merely encapsulate the ganglion cells. These autonomic ganglia link into longitudinal sympathetic trunks. Some of the neuroblasts migrate farther and assemble as collateral ganglia—ganglia not linked into longitudinal trunks. Still others migrate near, or within, the visceral organs that they will innervate and produce terminal ganglia. These ganglia are characteristic of the parasympathetic system.

Some cells of certain primitive collateral ganglia leave and invade the amassing mesodermal cortex of each adrenal gland. Consolidating in the centre, they become the endocrine cells of the medulla.

Sense organs. *Olfactory organ.* Paired thickenings of ectoderm near the tip of the head infold and produce olfactory pits. These expand into sacs in which only a relatively small area becomes olfactory in function (Figure 30E). Some epithelial cells in these regions remain as inert supporting elements. Others become spindle-shaped olfactory cells. One end of each olfactory cell projects receptive olfactory hairs beyond the free surface of the epithelium. From the other end a nerve fibre grows back and makes a connection within the brain.

Gustatory organ. Most taste buds arise on the tongue. Each bud, a barrel-shaped specialization within the epithelium that clothes certain lingual papillae (small projections on the tongue), is a cluster of tall cells, some of which have differentiated into taste cells whose free ends bear receptive gustatory hairs. Sensory nerve fibres end at the surface of such cells. Other tall cells are presumably inertly supportive in function.

Eye. The earliest indication of an eye is an optic vesicle (sac) bulging from each side of the forebrain (Figure 30B, C). It quickly becomes an indented optic cup, connected to the brain by a slender optic stalk. Most of the cup will become the retina, but its rim represents the epithelial part of the insensitive ciliary body and iris. The thicker inner layer of the cup becomes the neural layer of the retina, and by the sixth month three strata of neurons are recognizable in it: (1) visual cells, each bearing either a photoreceptive rod or a cone at one end; (2) bipolar cells, intermediate in position; (3) ganglion cells, which sprout axons that grow back through the optic stalk and make connections within the brain. The thin outer layer of the cup remains a simple epithelium whose cells gain pigment and make up the pigment epithelium of the retina.

The lens arises as a thickening of the ectoderm adjacent to the optic cup. It inpockets to form a lens vesicle and then detaches. The cells of its back wall become tall, transparent lens fibres. Mesoderm surrounding the optic cup specializes into two accessory coats. The outer coat, the tough, white sclera, is continuous with the transparent cornea. The inner coat, the vascular choroid, continues as the vascular and muscular ciliary body and the vascularized tissue of the iris. The eyelids are folds of adjacent skin; from the inside of each upper lid several lacrimal glands bud out (see SENSORY RECEPTION: *Human vision: Structure and function of the eye*).

Ear. The projecting part (auricle) of the external ear develops from hillocks on the first and second branchial arches (Figure 30C–E). The ectodermal groove between those arches deepens and becomes the external auditory canal. The auditory tube and tympanic cavity—the cavity at the inner side of the eardrum—are expansions of the endodermal pouch located between the first and second branchial arches. The area where ectodermal groove and endodermal pouch come in contact is the site of the future eardrum. The chain of three auditory ossicles (small bones) that stretches across the tympanic cavity is a derivative of the first and second arches.

Origins of the ear

The epithelium of the internal ear is at first a thickening of ectoderm at a level midway of the hindbrain (Figure 30B). This plate inpockets and pinches off as a closed sac, the otocyst. Its ventral part elongates and coils to resemble a snail's shell, thereby forming the cochlear duct, or seat of the organ of hearing. A middle region of the otocyst becomes chambers known as the utricle and saccule, related to the sense of balance. The dorsal part of the otocyst remodels drastically into three semicircular ducts, related to the sense of rotation. Fibres of the acoustic nerve grow among specialized receptive cells differentiated in certain regions of these three divisions.

MESODERMAL DERIVATIVES

Skeletal system. Except for part of the skull, all bones pass through three stages of development: membranous, cartilaginous, and osseous. The earliest ossification centres appear in the eighth week, but some do not arise until childhood years and even into adolescence.

Axial skeleton. The ventromedial walls (the walls toward the front and the midline) of the paired somites break down, and their cells migrate toward the axial notochord and surround it. Differentiation and growth of these segmental masses produce the jointed vertebrae. Ribs also grow out of each primitive vertebral mass, but they become long only in the thoracic region. Here their ventral ends join the sternum, which arises independently by the fusion of a pair of bars.

The skull has three components, different in origin. Its

basal region is an ancient heritage whose bones pass through the three typical stages of development. By contrast, the sides and roof of the skull develop directly from membranous primordia, or rudiments. The jaws are derivatives of the first pair of cartilaginous branchial arches but develop as membrane bone. Ventral ends of the second to fifth arches contribute the cartilages of the larynx and the hyoid bone (a bone of horseshoe shape at the base of the tongue). Dorsal ends of the first and second arches become the three auditory ossicles (the small bones in the middle ear).

Appendicular skeleton. The limb bones develop in three stages from axial condensations in the local mesoderm. The shoulder and pelvic supports are comparable sets, as are the bones of the arms and legs.

Articulations. Some type of joint exists wherever bones meet. Joints that allow little or no movement consist of connective tissue, cartilage, or bone. Movable joints arise as fluid-filled clefts in mesoderm, which condenses peripherally into a fibrous capsule.

Myoblast development

Muscular system. Much of each somite differentiates into myoblasts (primitive muscle cells) that become voluntary muscle fibres. Aggregations of such fibres become muscles of the neck and trunk. Muscles of the head and some of the neck muscles originate from mesoderm of branchial arches. Muscles of the limbs seemingly arise directly from local mesoderm. In general, muscle primordia may fuse into composites, split into subdivisions, or migrate away from their sites of origin. During these changes they retain their original nerve supply. Regardless of differences in source of origin, all voluntary muscle fibres are of the same striated type (marked by dark and light stripes). Spontaneous movements begin to occur in embryos about 10 weeks old. In general, involuntary muscle differentiates from mesoderm surrounding hollow organs; only the cardiac muscle type is striated.

Vascular system. All hollow organs, including arteries, veins, and lymphatics, are lined with epithelium—the principal functional tissue—and are ensheathed with muscular and fibrous coats.

Blood vessels. Primitive blood vessels arise in the mesoderm as tiny clefts bordered by flat endothelial cells. Growth and coalescence produce networks, out of which favoured channels persist as definite vessels, while others decline and disappear. A bilaterally symmetrical system of vessels is well represented in embryos four weeks old. This early plan is profoundly altered and made somewhat asymmetrical during the second month by fusions, atrophies, emergence of new vessels, and rerouting of older ones. The alterations reflect adjustments to changing form and pattern within the developing organ systems.

Arteries cranial to the heart (headward of the heart) are mostly products of the paired aortic arches, which course axially within the branchial arches, thus interconnecting the ventral aorta with paired dorsal aortas. The third pair of aortic arches becomes the common carotids; the fourth pair, the aortic arch and brachiocephalic; the fifth pair, the pulmonary arteries and ductus arteriosus. The dorsal aortas fuse into the single descending aorta, which bears three sets of paired, segmental branches. The dorsal set becomes the subclavian, intercostal, and lumbar arteries. The lateral set becomes arteries to the diaphragm, the adrenal glands, the kidneys, and the sex glands. The ventral set becomes the celiac, mesenteric, and umbilical arteries. Axial arteries to both sets of limb buds emerge from an original plexus, but they undergo drastic alteration and extensive replacement.

The primitive veins are symmetrically bilateral. They consist of vitelline veins from the yolk sac, umbilical veins from the placenta, and precardinal and postcardinal veins from the cranial and caudal regions (the regions toward the head and toward the tail) of the body. Drastic transformations occur in all of these, and new pairs of veins (subcardinals and supracardinals) arise also, caudal to the heart. From the vitellines come chiefly the portal and hepatic veins. The left umbilical becomes the main return from the placenta by making a diagonal channel, the ductus venosus, through the liver to the heart. The precardinal veins change their names to the internal jugulars, but near the heart an interconnection permits both to drain into a common stem, then called the superior vena cava. Caudal to the heart, the postcardinals virtually disappear, and all blood return shifts to the right side as a new compound vessel, the inferior vena cava, becomes dominant. Pulmonary veins open into the left atrium. Veins from the limb buds organize from an early peripheral border vein.

Lymphatic vessels. The lymph vessels develop independently in close association with veins. Linkages produce the thoracic duct, which is the main drainage return for lymph. Masses of lymphocytes accumulate about lymphatic vessels and organize as lymph nodes. The spleen has somewhat similar tissue, but its channels are supplied with blood.

Origins of the heart

Heart. Fusion combines two endothelial tubes, and these are surrounded by a mantle of mesoderm that will become the muscular and fibrous coats of the heart. At three weeks the heart is a straight tube that is beginning to beat (Figure 29M). Starting at the head end, four regions can be recognized: bulbus, ventricle, atrium, and sinus venosus. Since the heart is anchored at both ends, rapid elongation forces it to bend. In doing this, the sinus–atrium and bulbus–ventricle reverse their original relations. Further development concerns the transformation of a single-chambered heart into one with four chambers.

The atrium becomes subdivided by the growth of two incomplete partitions, or septa, placed close together and each covering the defect in the other. The ventricle also subdivides, but by a single, complete partition. A canal, connecting atria and ventricles, becomes two canals. The bulbus is absorbed into the right ventricle, and its continuation (the truncus) subdivides lengthwise, forming the aorta and the pulmonary artery. The right horn of the sinus venosus is absorbed into the right atrium, together with the superior and inferior venae cavae, which originally drained into the sinus. The transverse portion of the sinus persists as the coronary sinus. The pulmonary veins retain their early drainage into the left atrium. Important valves develop and ensure flow within the heart from atria to ventricles, and outward from the ventricles into the aorta and the pulmonary artery.

Birth initiates breathing, and the abandonment of the placental circulation follows. These changes entail a drastic rerouting of blood through the heart. As a result, the two atrial septa fuse and no longer permit blood to pass from the right atrium to the left atrium. Blood in the pulmonary artery no longer virtually bypasses the lungs; previously it had passed to the aorta directly through a shunt offered by the ductus arteriosus. As a sequel to these changes, the abandoned umbilical arteries, umbilical vein, ductus venosus, and ductus arteriosus all collapse and become fibrous cords.

Origins of the kidney, urinary bladder, and urethra

Urinary system. Vertebrates have made three experiments in kidney production: the pronephros, or earliest type; the mesonephros, or intermediate kidney; and the metanephros, or permanent kidney. All arise from the cellular plates called nephrotomes that connect somites with the mesodermal sheets that bound the body cavity. The vestigial pronephros is represented solely by several pairs of tubules; they join separately formed excretory ducts that grow downward and enter the cloaca, the common outlet for urine, genital products, and for intestinal wastes. Next tailward arise some 40 pairs of nephric (kidney) tubules that constitute the mesonephros; these tubules join the same excretory ducts, hereafter called the mesonephric ducts. The two sets of mesonephric tubules serve as functioning kidneys until the 10th week.

Each permanent kidney, or metanephros, develops still farther tailward. A so-called ureteric primordium buds off each mesonephric duct, near its hind end. The ureteric stem elongates and expands terminally, thereby forming the renal pelvis and calices; continued bushlike branching produces collecting ducts. The early ureteric bud invades a mass of nephrotome tissue. The branching collecting ducts progressively break this tissue up into tiny lumps, each of which becomes a long secretory tubule, or nephron, and joins a nearby terminal twig of the duct system. Continued proliferation of ducts and nephric tissue produces over a million urine-producing tubules in each kidney.

The blind caudal end of the endodermal hindgut absorbs the stem of each mesonephric duct, whereupon the remainder of the duct and the ureter acquire separate openings into the hindgut. This expanded region of the gut, now a potential receptacle for feces, urine, and reproductive products, is known as a cloaca. It next subdivides into a rectum behind and a urogenital sinus in front. The sinus, in turn, will specialize into the urinary bladder and the urethra. The prostate gland develops as multiple buds from the urethra, close to the bladder.

Origins of sex glands and ducts and external genitalia

Genital system. The genital organs begin to develop in the second month, but for a time sex is not grossly distinguishable. Also, a double set of male and female ducts arise, and not until later does the unneeded set decline. Hence, this period is commonly called the indifferent stage.

Gonads. Sex glands develop in a pair of longitudinal ridges located alongside the mesentery, the anchoring fold of membrane to the gut. The primordial sex cells appear first in the cloacal wall, from which they migrate upward in the gut, pass through its mesentery, and finally invade the genital ridges, where they proliferate. The testes are the earliest type of gonad to organize. They begin by developing testis cords and a testis capsule. The cords radiate from one focal point at the periphery, and thin fibrous partitions segregate groups of the cords within wedge-shaped compartments. These cords do not gain channels and become semen-producing tubules until near the time of puberty. The ovaries organize somewhat tardily by differentiating an outer portion, the cortex, and a central portion, the medulla. The cortex contains the primordial sex cells; these become surrounded by a layer of ordinary cells, thereby forming primary ovarian follicles. Both the testes and the ovaries undergo relative shifts from their early sites to lower positions in the body. But only the testes make a bodily descent; this is into the scrotum.

Genital ducts. In the male, a few mesonephric tubules on each side do not degenerate but link up with the neighbouring testis tubules. The converted mesonephric tubules and the retained mesonephric ducts become the male sex ducts. Near their terminations they outpouch seminal vesicles and then open into the urethra. In the female, a pair of ducts develops from the epithelium clothing the mesonephric ridges. These ducts, known as the uterine tubes, mostly parallel the courses of the mesonephric ducts, but at their lower ends they unite into a common tube that becomes the uterus and vagina.

External genitalia. Both sexes develop a genital tubercle (*i.e.*, a knob) and a pair of urogenital folds flanked by a pair of genital swellings. At three months these rudiments begin to assume male or female characteristics. In the male, the tubercle and the united urogenital folds combine as the penis, thereby continuing the urethra to its end; the genital swellings shift toward the anus, fuse, and become the scrotum. In the female, the tubercle remains small, as the clitoris; it does not contain the urethra. The urogenital folds remain unclosed as the lesser vulvar lips and are flanked by the unshifted and unfused genital swellings, or greater lips.

Origins of the body cavities

Coelom. The lateral mesoderm, beyond the somites and nephrotomes, splits into two layers: the somatic layer and, underlying the somatic layer, the splanchnic layer. The intervening space is the coelom. As the embryo's body folds off, its coelom becomes a single closed cavity. In it can be recognized, regionally, a provisional pericardial cavity (cavity for the heart), two pleural canals (for the lungs), and a peritoneal cavity (for the abdominal contents). A thick plate of mesoderm, the transverse septum, constitutes a partial partition just ahead of the developing liver. Two pairs of membranes grow out from the septum. One set separates the pericardial cavity from the two pleural cavities; these membranes later expand into the pericardium and enclose the heart. The other pair of membranes separates the pleural cavities from the peritoneal cavity of the abdomen. The definitive diaphragm is a composite partition, much of which is furnished by the transverse septum; lesser contributions are from the lateral body walls and the paired membranes that separated the pleural and peritoneal cavities.

ENDODERMAL DERIVATIVES

Pharynx. The tongue is a product of four branchial arches, whose ventral ends merge in its midplane. Papillae elevate from the surface, and taste buds arise as specializations within the covering epithelium of some of them. Pharyngeal pouches are early lateral expansions of the local endoderm, alternating with the branchial arches. The first pair elongate as the auditory tubes and tympanic cavities. The second pair mark the site of the tonsils. The third pair give rise to the halves of the thymus, and the third and fourth pairs produce the two sets of parathyroid glands. The thyroid gland buds off the pharyngeal floor in the midplane and at the level of the second branchial arches.

Development of the esophagus, stomach, and intestines

Digestive tube. As the embryo folds off, the endoderm is rolled in as the foregut and hindgut. Continued growth progressively closes both the midbody and the midgut (Figure 29P, Q). The esophagus remains as a simple, straight tube. The stomach grows faster on its dorsal side, thereby forming the bulging greater curvature; the stomach also rotates 90° so that its original dorsal and ventral borders come to lie left and right. The intestine elongates faster than the trunk, so that its loops find temporary room by pushing into the umbilical cord. Later, the loops return, completing a rotation that gives the characteristic final placement of the small and large intestines.

When the gut folds into a tube it is suspended by a sheetlike dorsal mesentery, or membranous fold. In the region of the stomach it forms an expansive pouch, the omental bursa. Secondary fusions of the bursa and of some of the rest of the mesentery with the body wall produce lines of attachment from stomach to rectum inclusive, different from the original midplane course. Such fusions also anchor firmly some parts of the tract. A ventral mesentery, beneath the gut, exists only in the region of the stomach and liver.

Major glands. The liver arises as a ventral outgrowth of the foregut that invades the early transverse septum. Although rapid growth causes it to bulge prominently away from this septum, it remains attached to the septum and hence to the definitive diaphragm. The differentiating glandular tissue takes the form of plates bathed by blood channels. The stem of the original liver bud becomes the common bile duct, whereas a secondary outgrowth produces the cystic duct and the gallbladder.

The pancreas takes its origin from a larger dorsal bud and a smaller ventral bud, both off the foregut. The two merge and their ducts communicate, but in man it is the lesser, ventral duct that becomes the stem outlet. Secretory acini are berrylike endings of the branching ducts. Pancreatic islets arise as special sprouts from the ducts; these differentiate into endocrine tissue that secretes insulin.

Nose, larynx, trachea, and lungs

Respiratory system. *Nasal cavity.* The first part of the respiratory system is ectodermal in origin. The olfactory sacs (see above *Ectodermal derivatives: Sense organs*) become continuous secondarily with a passage captured from the primitive mouth cavity. This addition is produced by a horizontal partition, the palate. It arises from a pair of shelflike folds that grow out from the halves of the primitive upper jaw and then unite. The final nasal passage extends from the nostrils to the back of the pharynx.

Larynx, trachea, and lungs. A hollow lung bud grows off the floor of the endodermal pharynx, just caudal (tailward) to the pharyngeal pouches and in the midline. It has the form of a tube with an expanded end. The entrance to this tube is the glottis, and the region about it becomes the larynx. The tube proper represents the trachea. Its terminal expansion divides into two branches, and these tubes elongate as the primary bronchi. Continued growth and budding produce two side branches from the right bronchus and one from the left. These branches and the blind ends of the two parent bronchi indicate the future plan of the lungs, with three right lobes and two left lobes. Continued branchings, through the sixth month, produce bronchioles of different orders. In the final months the smaller ducts and early respiratory alveoli (air sacs) appear, the lungs losing their previous glandular appearance and also becoming highly vascular. Until breathing distends the lungs, these organs remain relatively small.

(L.B.A.)

Postnatal development

Human growth is far from being a simple and uniform process of becoming taller or larger. As a child gets bigger, there are changes in shape and in tissue composition and distribution. In the newborn infant the head represents about a quarter of the total length; in the adult it represents about one-seventh. In the newborn infant the muscles constitute a much smaller percentage of the total body mass than in the young adult. In most tissues, growth consists both of the formation of new cells and the packing in of more protein or other material into cells already present; early in development cell division predominates and later cell filling.

TYPES AND RATES OF HUMAN GROWTH

'he omplex atterns of rowth

Different tissues and different regions of the body mature at different rates, and the growth and development of a child consists of a highly complex series of changes. It is like the weaving of a cloth whose pattern never repeats itself. The underlying threads, each coming off its reel at its own rhythm, interact with one another continuously, in a manner always highly regulated and controlled. The fundamental questions of growth relate to these processes of regulation, to the program that controls the loom, a subject as yet little understood. Meanwhile, height is in most circumstances the best single index of growth, being a measure of a single tissue (that of the skeleton; weight is a mixture of all tissues, and this makes it a less useful parameter in a long-term following of a child's growth). In this section, the height curves of girls and boys are considered in the three chief phases of growth; that is (briefly) from conception to birth, from birth until puberty, and during puberty. Also described are the ways in which other organs and tissues, such as fat, lymphoid tissue, and the brain, differ from height in their growth curves. There is a brief discussion of some of the problems that beset the investigator in gathering and analyzing data about growth of children, of the genetic and environmental factors that affect rate of growth and final size, and of the way hormones act at the various phases of the growth process. Lastly, there is a brief look at disorders of growth. Throughout, the emphasis is on ways in which individuals differ in their rates of growth and development.

rowth as istance nd elocity

The changes in height of the developing child can be thought of in two different ways: the height attained at successive ages and the increments in height from one age to the next, expressed as rate of growth per year. If growth is thought of as a form of motion, the height attained at successive ages can be considered the distance travelled, and the rate of growth, the velocity. The velocity or rate of growth reflects the child's state at any particular time better than does the height attained, which depends largely on how much the child has grown in all preceding years. The blood and tissue concentrations of those substances whose amounts change with age are thus more likely to run parallel to the velocity rather than to the distance curve. In some circumstances, indeed, it is the acceleration rather than the velocity curve that best reflects physiological events.

In general, the velocity of growth decreases from birth onward (and actually from as early as the fourth month of fetal life; see below), but this decrease is interrupted shortly before the end of the growth period. At this time, in boys from about 13 to 15 years (Figure 33), there is marked acceleration of growth, called the adolescent growth spurt. From birth until age four or five, the rate of growth in height declines rapidly, and then the decline, or deceleration, gets gradually less, so that in some children the velocity is practically constant from five or six up to the beginning of the adolescent spurt. A slight increase in velocity is sometimes said to occur between about six and eight years.

This general velocity curve of growth in height begins a considerable time before birth. Figure 31 shows the distance and velocity curves for body length in the prenatal period and first two postnatal years. The peak velocity of length is reached at about four months after the mother's last menstruation. (Age in the fetal period is usually reckoned from the first day of the last menstrual period, an average of two weeks before actual fertilization, but, as a rule, the only locatable landmark.)

Growth in weight of the fetus follows the same general pattern as growth in length, except that the peak velocity is reached much later, at approximately 34 weeks after the mother's last menstrual period.

From D. Thompson, *Growth and Form*, 2nd ed. (1963); Cambridge University Press

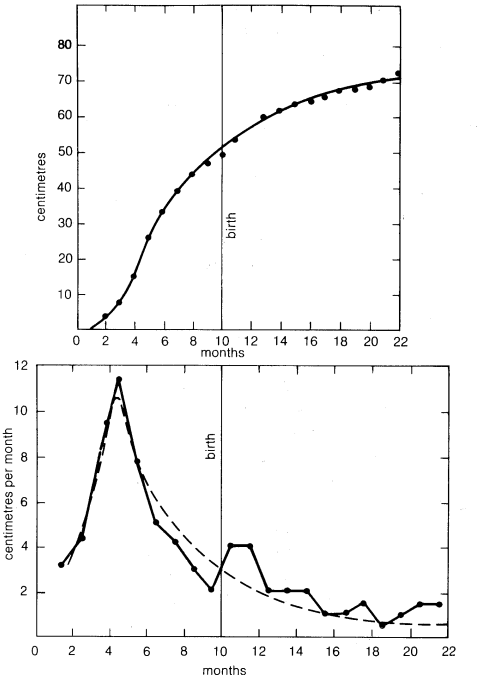


Figure 31: Curves indicating (above) growth attained and (below) velocity of growth during prenatal and early postnatal periods. The dashed line in the lower chart represents a smoothing out of the curve between points.

There is considerable evidence that from about 34 to 36 weeks onward the rate of growth of the fetus slows down because of the influence of the maternal uterus, whose available space is by then becoming fully occupied. Twins slow down earlier, when their combined weight is approximately the 36-week weight of a single fetus. Babies who are held back in this way grow rapidly as soon as they have emerged from the uterus. Thus there is a significant negative association between weight of a baby at birth and weight increment during the first year; in general, larger babies grow less, the smaller more. For the same reason there is practically no relation between adult size and the size of that person at birth, but a considerable relation has developed by the time the person is two years old. This slowing-down mechanism enables a genetically large child developing in the uterus of a small mother to be delivered successfully. It operates in many species of animals; the most dramatic demonstration was by crossing reciprocally a large Shire horse and a small Shetland pony. The pair in which the mother was a Shire had a large newborn foal, and the pair in which the mother was Shetland had a small foal. But both foals were the same size after a few months, and when fully grown both were about halfway between their parents. The same has been shown in cattle crosses.

Rapid growth following small size at birth

Poor environmental circumstances, especially of nutrition, result in lowered birth weight in the human being. This seems chiefly to be caused by a reduced rate of growth in the last two to four weeks of fetal life, for weights of babies born in 36 or 38 weeks in various parts of the world in various circumstances are said to be similar. Mothers who, because of adverse circumstances in their own childhood, have not achieved their full growth potential may produce smaller fetuses than they would have, had they grown up in better circumstances. Thus two generations or even more may be needed to undo the effect of poor environmental circumstances on birth weight.

The great rate of growth of the fetus compared with that of the child is largely due to the fact that cells are still multiplying. The proportion of cells undergoing mitosis (the ordinary process of cell multiplication by splitting) in any tissue becomes progressively less as the fetus gets older, and it is generally thought that few if any new nerve cells (apart from the cells in the supporting tissue, or neuroglia) and only a limited proportion of new muscle cells appear after six postmenstrual months, the time when the velocity in linear dimensions is dropping sharply.

Prenatal and postnatal growth contrasted

The muscle and nerve cells of the fetus are considerably different in appearance from those of the child or adult. Both have little cytoplasm (cell substance) around the nucleus. In the muscle there is a great amount of intercellular substance and a much higher proportion of water than in mature muscle. The later fetal and the postnatal growth of the muscle consists chiefly of building up the cytoplasm of the muscle cells; salts are incorporated and the contractile proteins formed. The cells become bigger, the intercellular substance largely disappears, and the concentration of water decreases. This process continues quite actively up to about three years of age and slowly thereafter; at adolescence it briefly speeds up again, particularly in boys, under the influence of androgenic (male sex) hormones. In the nerve cells cytoplasm is added and elaborated, and extensions grow that carry impulses from and to the cells—the axons and dendrites, respectively. Thus postnatal growth, for at least some tissues, is chiefly a period of development and enlargement of existing cells, while early fetal life is a period of division and addition of new cells.

TYPES OF GROWTH DATA

Growth is in general a regular process. Contrary to what is said in some of the older textbooks, growth in height does not proceed by fits and starts, nor does growth in upward dimensions alternate with growth in transverse ones. The more carefully measurements are taken, with precautions, for example, to minimize the decrease in height that occurs during the day for postural reasons, the more regular does the succession of points in a graph of growth become. Many attempts have been made at finding mathematical curves that fit, and thus summarize, human growth data. What is needed is a curve or curves with relatively few constants, each capable of being interpreted in a biologically meaningful way. Yet the fit to empirical data must be adequate within the limits of measuring error. The problem is difficult, partly because the measurements usually taken are themselves biologically complex. Stature, for example, consists of leg length and trunk length and head height, all of which have rather different growth curves. Even with relatively homogeneous dimensions such as the length of the radius bone in the forearm, or width of an arm muscle, it is not clear what purely biological assumptions should be made as the basis for the form of the curve. The assumption that cells are continuously dividing leads to a different formulation from the assumption that cells are adding constant amounts of nondividing material or amounts of nondividing material at rates varying from one age period to another.

Fitting a curve to the individual values, however, is the only way of extracting the maximum information from an individual's measurement data. More than one curve is needed to fit the postnatal age range. It seems that two curves may suffice, at least for many measurements such as height and weight—one curve for the period from a few months after birth to the beginning of adolescence and a different type of curve for the adolescent spurt.

Such curves have to be fitted to data on single individuals. Yearly averages derived from different children each measured only once do not, in general, give the same curve. Thus the distinction between the two sorts of investigation is important. When the same child at each age is used, the study is called longitudinal; when different children at each age are used, it is called cross-sectional. In a cross-sectional study all of the children at age eight, for example, are different from those at age seven. A study may be longitudinal over any number of years; there are short-term longitudinal studies extending from age four to six, for instance, and full birth-to-maturity longitudinal studies in which the children may be examined once, twice, or more times every year from birth until 20 or over. Mixed longitudinal studies are those in which children join and leave the group studied at varying intervals. Both cross-sectional and longitudinal studies have their uses, but they do not give the same information, and the same statistical methods cannot be used for the two types of study. Cross-sectional surveys are obviously cheaper and more quickly done and can include much larger numbers of children. Periodic cross-sectional surveys are valuable in assessing the nutritional progress of a country or a socioeconomic group and the health of the child population as a whole. But they never reveal individual differences in rate of growth or in the timing of particular phases such as the adolescent growth spurt. It is these individual rate differences that throw light on the genetic control of growth and on the correlation of growth with psychological development, educational achievement, and social behaviour.

Advantages in longitudinal and cross-sectional studies

Longitudinal studies are laborious and time-consuming; they demand great perseverance on the part of those who make them and those who take part in them; and they demand high technical standards, since in the calculation of a growth increment from one occasion to the next opportunities for two errors of measurement occur. In spite of these problems, longitudinal studies are the indispensable base on which the diagnosis and treatment of disorders of growth rest, for the clinical approach is a longitudinal one; and each child treated with human growth hormone, or with other hormones that affect growth, represents an attempt to alter an individual pattern of growth velocity.

Averages simply computed from cross-sectional data inevitably produce velocity curves that are flatter and broader than the curve for an individual and hence not a proper basis for clinical standards. It is possible to construct curves, however, whose 50th percentile (or average) represents the actual growth of a typical individual, by taking the shape of the curve from individual longitudinal data and the absolute values for the beginning and end from large cross-sectional surveys. Figures 32 and 33 show height-attained and height-velocity curves for the "typical" boy and girl in Britain in 1965, determined in this way. By "typical" is meant that boy or girl who has the mean (average) birth length, grows always at the mean velocity, has the peak of the adolescent growth spurt at the mean age, and finally reaches the mean adult height at the mean age of cessation of growth. Practically no individual follows the 50th percentile curve of Figure 32, but most have curves of the same shape. Standards for height for clinical use are constructed around these curves.

From *Archives of Disease in Childhood* (1966)

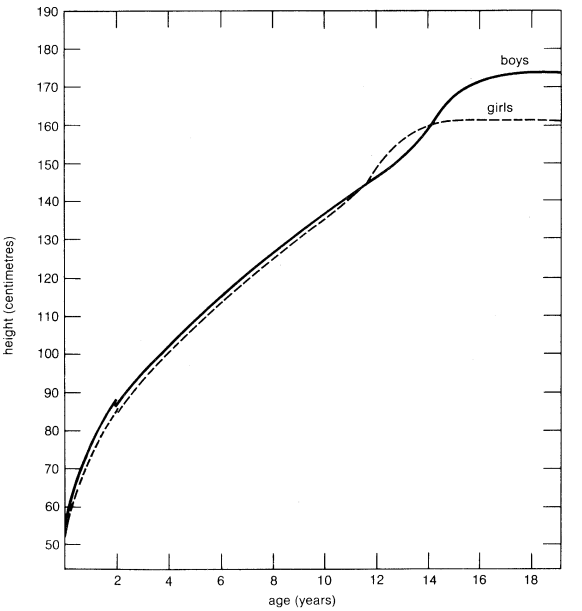


Figure 32: Individual height curves for British boys and girls in 1965 (see text).

Measuring the child's height

Boys' and girls' height curves. Figures 32 and 33 show the height curves from birth to maturity. Up to age two, the child is measured lying on his back. One examiner holds his head in contact with a fixed board, and a second person stretches him out to his maximum length and then brings a moving board into contact with his heels. This measurement, called supine length, averages about one centimetre more than the measurement of standing height taken on the same child, hence the break in the line in Figure 32 at age two. This occurs even when, as in the best techniques, the child is urged to stretch upwards to the full and is aided in doing so by a measurer's applying gentle upward pressure to his mastoid processes.

Figure 32 shows the typical girl as slightly shorter than the typical boy at all ages until adolescence. She becomes taller shortly after age 11 because her adolescent spurt takes place two years earlier than the boy's. At age 14 she is surpassed again in height by the typical boy, whose adolescent spurt has now started, while hers is nearly finished. In the same way, the typical girl weighs a little less than the boy at birth, equals him at age eight, becomes heavier at age nine or 10, and remains so until about age 14½.

The velocity curves given in Figure 33 show these processes more clearly. At birth the typical boy is growing slightly faster than the typical girl, but the velocities become equal at about seven months, and then the girl grows

From *Archives of Disease in Childhood*, vol. 41, pp. 454, 613 (1966)

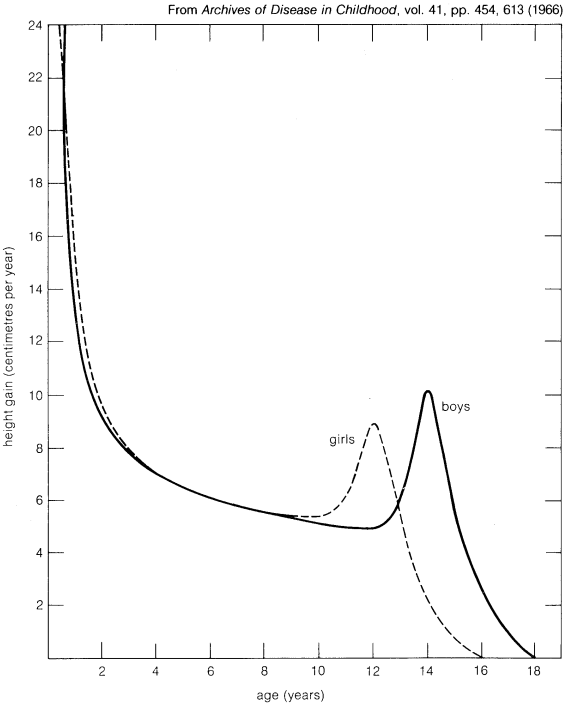


Figure 33: Typical velocity curves for supine length or height in British boys and girls in 1965 (see text).

faster until four years. From then until adolescence no differences in velocity can be detected. The sex difference is best thought of, perhaps, in terms of acceleration, the boy decelerating harder than the girl over the first four years.

Different tissues and parts of the body. The majority of skeletal and muscular dimensions follow approximately the growth curve described for height, and so also do the dimensions of the internal organs such as the liver, the spleen, and the kidneys. But some exceptions exist, most notably the brain and skull, the reproductive organs, the lymphoid tissue of the tonsils, adenoids, and intestines, and the subcutaneous fat.

In Figure 34 these differences are shown; the size attained by various tissues is given as a percentage of the birth-to-maturity increment. Height follows the "general" curve. The reproductive organs, internal and external, have a slow prepubescent growth, followed by a large adolescent spurt; they are less sensitive than the skeleton to one set of hormones and more sensitive to another.

The brain, together with the skull covering it and the eyes and ears, develops earlier than any other part of the body and thus has a characteristic postnatal curve. At birth it is already 25 percent of its adult weight, at age five about 90 percent, and at age 10 about 95 percent. Thus if the brain has any adolescent spurt at all, it is a small one. A small but definite spurt occurs in head length and breadth, but all or most of this is due to thickening of the skull bones and the scalp, together with development of the air sinuses.

Growth of the brain, the skull, and the face

The dimensions of the face follow a path somewhat closer to the general curve. There is a considerable adolescent spurt, especially in the lower jaw, or mandible, resulting in the jaw's becoming longer and more projecting, the profile straighter, and the chin more pointed. As always

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications, after Scammon

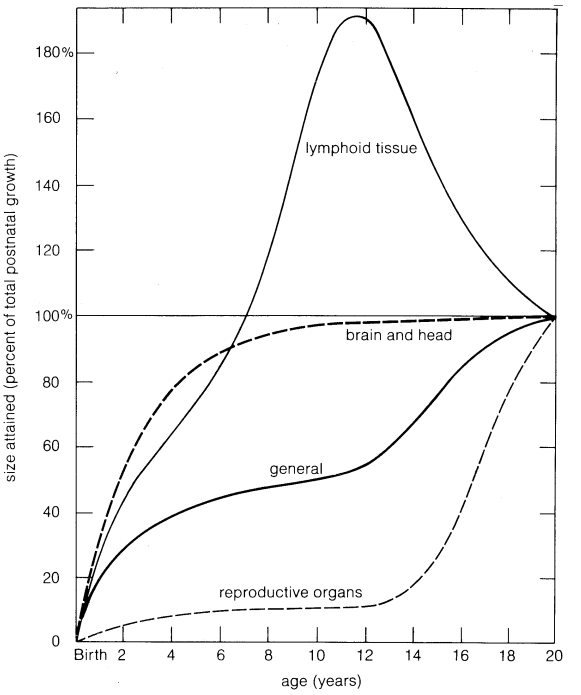


Figure 34: Major types of postnatal growth of parts and organs of the body (see text).

in growth, there are considerable individual differences, to the point that a few children have no detectable spurt at all in some face measurements.

The eye probably has a slight adolescent spurt, which is probably responsible for the increase in frequency of shortsightedness in children that occurs at the time of puberty. Though the degree of myopia increases continuously from at least age six to maturity, a particularly rapid rate of

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

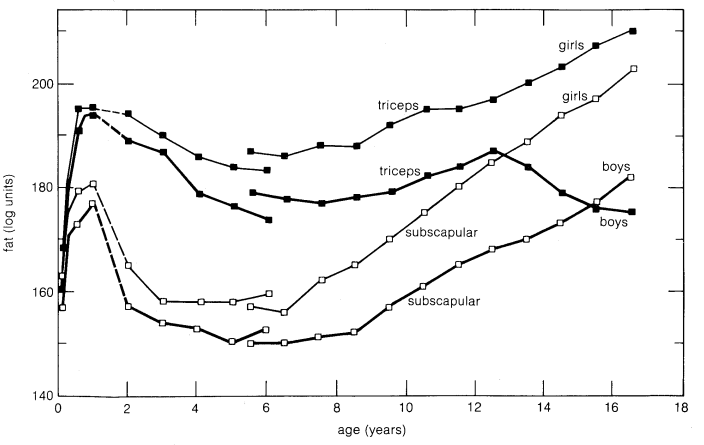


Figure 35: Amount of subcutaneous fat on the back of the arm (triceps) and the back of the body (subscapular) from birth to 16 years (see text).

change occurs at about 11 to 12 in girls and 13 to 14 in boys, and this would be expected if there was a rather greater spurt in the axial dimension (the dimension from front to back) of the eye than in its vertical dimension.

The lymphoid tissue has quite a different growth curve from the rest. It reaches its maximum amount before adolescence and then, probably under the direct influence of sex hormones, declines to its adult value.

Development of fat

The subcutaneous fat layer also has a curve of its own, of a slightly complicated sort. Its thickness can be measured either by X-rays or, more simply, at certain sites in the body, by picking up a fold of skin and fat between the thumb and forefinger and measuring its thickness with a special, constant-pressure caliper. Figure 35 shows the distance curves of skin folds taken halfway down the back of the upper arm (the triceps) and on the body just below the angle of the shoulder blade, or scapula. Subcutaneous fat begins to be laid down in the fetus at about 34 weeks postmenstrual age, increases from then until birth and from birth onward until about nine months. (This is in the average child; the peak may be reached as early as six months or as late as 12 or 15.) After nine months, when the velocity of fat gain is zero, the fat usually decreases (that is, it has a negative velocity) until age six to eight, when it begins to increase once more. Girls have a little more fat than boys at birth, and the difference becomes more marked during the period of loss, since girls lose less than boys. From eight years on, the curves for girls and boys diverge more radically, as do the curves for limb and body fat. At adolescence the limb fat in boys decreases, while the body fat shows a temporary slowing down of gain but no actual loss. In girls there is a slight halting of the limb-fat gain at adolescence, but no loss; the trunk fat shows only a steady rise until adolescence.

Development at puberty

ALTERATIONS IN GROWTH RATE

At puberty, a considerable alteration in growth rate occurs. There is a swift increase in body size, a change in shape and composition of the body, and a rapid development of the gonads, or sex glands—the reproductive organs and the characters signalling sexual maturity. Some of these changes are common to both sexes, but most are sex-specific. Boys have a great increase in muscle size and strength, together with a series of physiological changes making them capable of doing heavier physical work than girls and of running faster and longer. These changes all specifically adapt the male to his primitive primate role of dominating, fighting, and foraging. Such adolescent changes occur generally in primates (that is, men, apes, and monkeys) but are more marked in some species than in others. Man lies at about the middle of the primate range, as regards both adolescent size increase and degree of sexual differentiation.

During the adolescent spurt in height, for a year or more, the velocity of growth approximately doubles; a boy is likely to be growing again at the rate he last experienced about age two. The peak velocity of height (P.H.V., a point much used in growth studies) averages about 10.5 centimetres per year in boys and 9.0 centimetres in girls (about 4 and 3.4 inches, respectively), but this is the "instantaneous" peak given by a smooth curve drawn through the observations. The velocity over the whole year encompassing the six months before and after the peak is naturally somewhat less. During this year a boy usually grows between 7 and 12 centimetres (2.75 and 4.75 inches) and a girl between 6 and 11 centimetres (2.35 and 4.35 inches). Children who have their peak early reach a somewhat higher peak than those who have it late.

Age of peak height

The average age at which the peak is reached depends on the nature and circumstances of the group studied more, probably, than does the height of the peak. In moderately well-off British or North American children at present the peak occurs on average at about 14.0 years in boys and 12.0 years in girls. Though the absolute average ages differ from population to population, the two-year sex difference always persists.

Practically all skeletal and muscular dimensions take part in the spurt, though not to an equal degree. Most of the spurt in height is due to acceleration of trunk length rather than of length of legs. There is a fairly regular order in which the dimensions accelerate; leg length as a rule reaches its peak first, followed by the body breadths, with shoulder width last. The earliest structures to reach their adult status are the head, hands, and feet.

The spurt in muscle, of both limbs and heart, coincides with the spurt in skeletal growth, for both are caused by the same hormones. Boys' muscle widths reach a peak velocity of growth that is greater than that reached by girls. But, since girls have their spurt earlier, there is actually a period, from about 12½ to 13½, when girls on average have larger muscles than boys of the same age, as well as being taller (Figure 32). Simultaneously with the spurt there is a loss of fat, as described above.

The marked increase in muscle size in boys at adolescence leads to an increase in strength, illustrated in Figure 36. Before adolescence, boys and girls are similar in strength for a given body size and shape; after, boys

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

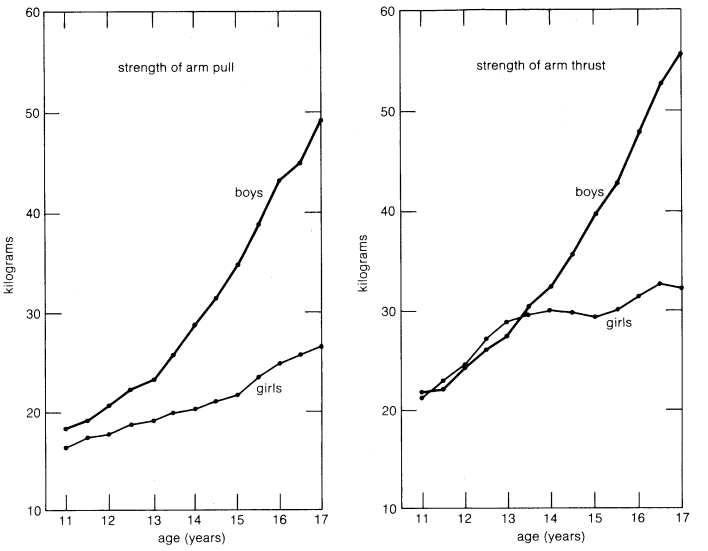


Figure 36: Strength of arm pull and thrust for boys and girls.

have much greater strength, probably due to development of more force per gram of muscle as well as to absolutely larger muscles. They also develop larger hearts and lungs relative to their size, a higher systolic blood pressure (the pressure resulting from a heart contraction), a lower resting heart rate, a greater capacity for carrying oxygen in the blood with more hemoglobin, and a greater power for neutralizing the chemical products of muscular exercise such as lactic acid. In short, the male becomes at adolescence more adapted for the tasks of hunting, fighting, and manipulating all sorts of heavy objects, as is necessary in some forms of food gathering.

Adolescent increase in strength

It is as a direct result of these anatomical and physiological changes that athletic ability increases so much in boys at adolescence. The popular notion of a boy's "outgrowing his strength" at this time has little scientific support. It is true that the peak velocity of strength is reached a year or so later than that of height, so that a short period may exist when the adolescent, having completed his skeletal and probably also his muscular growth, still does not have the strength of a young adult of the same body size and shape. But this is a temporary phase; considered absolutely, power, athletic skill, and physical endurance all increase progressively and rapidly throughout adolescence.

The adolescent spurt in skeletal and muscular dimensions is closely related to the rapid development of the reproductive system that takes place at this time. The course of this development is outlined diagrammatically in Figure 37. The bar marked "breast" in the chart for the girls and the bars marked "penis" and "testis" in the chart for the boys represent the periods of accelerated growth of these organs. Other bars indicate the genitalia rating and the advent and development of the pubic hair. The sequence

Development of reproductive organs

and timings that are given represent in each case average values for British boys and girls; the North American average is within two or three months of this. To give an idea of the individual departures from the average, figures for the range of age at which the various events begin and end are inserted under the first and last point of the bars. The acceleration of penis growth, for example, begins on average at about age 12½ years, but sometimes as early as 10½ and sometimes as late as 14½. The completion of penis development usually occurs at about age 14½, but in some boys is at 12½ and in others at 16½. There are a few boys, it will be noticed, who do not begin their spurts in height or penis development until the earliest

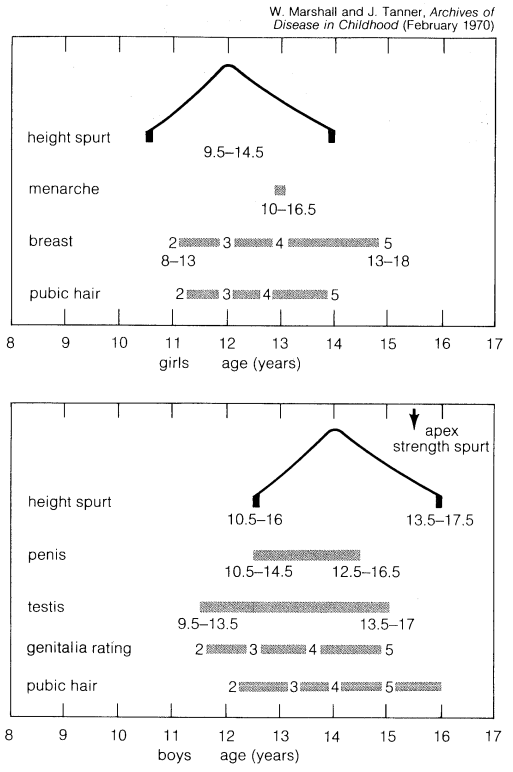


Figure 37: Sequence of events at adolescence for average British boys and girls. Range of ages within which each event may begin and end is given by the figures placed directly below its start and finish. The numbers 2 to 5 within the bars indicate stage of maturity, 5 being full maturity and 2 the beginning of adolescence (see text).

maturers have entirely completed theirs. At ages 13, 14, and 15 there is an enormous variability among any group of boys, who range all the way from practically complete maturity to absolute preadolescence. The same is true of girls aged 11, 12, and 13.

The psychological and social importance of this difference in the tempo of development, as it has been called, is great, particularly in boys. Boys who are advanced in development are likely to dominate their contemporaries in athletic achievement and sexual interest alike. Conversely the late developer is the one who all too often loses out in the rough and tumble of the adolescent world, and he may begin to wonder whether he will ever develop his body properly or be as well endowed sexually as those others whom he has seen developing around him. An important part of the educationist's and the doctor's task at this time is to provide information about growth and its variability to preadolescents and adolescents and to give sympathetic support and reassurance to those who need it.

The first sign of puberty in the boy

The sequence of events, though not exactly the same for each boy or girl, is much less variable than the age at which the events occur. The first sign of puberty in the boy is usually an acceleration of the growth of the testes and scrotum with reddening and wrinkling of the scrotal skin. Slight growth of pubic hair may begin about the same time but is usually a trifle later. The spurts in height and penis growth begin on average about a year after the first testicular acceleration. Concomitantly with the growth of the penis, and under the same stimulus, the seminal vesicles, the prostate, and the bulbo-urethral glands, all of which contribute their secretions to the seminal fluid, enlarge and develop. The time of the first ejaculation of seminal fluid is to some extent culturally as well as biologically determined but as a rule is during adolescence and about a year after the beginning of accelerated penis growth.

Axillary (armpit) hair appears on average some two years after the beginning of pubic hair growth; that is, when pubic hair is reaching stage 4 (Figure 37). There is enough variability and dissociation in these events, so that a very few children's axillary hair actually appears first. In boys, facial hair begins to grow at about the time that the axillary hair appears. There is a definite order in which the hairs of moustache and beard appear: first at the corners of the upper lip, then over all the upper lip, then at the upper part of the cheeks, in the midline below the lower lip, and, finally, along the sides and lower borders of the chin. The remainder of the body hair appears from about the time of first axillary hair development until a considerable time after puberty. The ultimate amount of body hair that an individual develops seems to depend largely on heredity, though whether because of the kinds and amounts of hormones secreted or because of variations in the reactivity of the end organs is not known.

Breaking of the voice occurs relatively late in adolescence. The change in pitch accompanies enlargement of the larynx and lengthening of the vocal cords, caused by the action of the male hormone testosterone on the laryngeal cartilages. There is also a change in quality that distinguishes the voice (more particularly the vowel sounds) of both male and female adults from that of children. This is caused by the enlargement of the resonating spaces above the larynx, as a result of the rapid growth of the mouth, nose, and maxilla (upper jaw).

In the skin, particularly of the armpits and the genital and anal regions, the sebaceous and apocrine sweat glands develop rapidly during puberty and give rise to a characteristic odour; the changes occur in both sexes but are more marked in the male. Enlargement of the pores at the root of the nose and the appearance of comedones (blackheads) and acne, while likely to occur in either sex, are considerably more common in adolescent boys than girls, since the underlying skin changes are the result of androgenic (male sex hormone) activity.

During adolescence the male breast undergoes changes, some temporary and some permanent. The diameter of the areola, which is equal in both sexes before puberty, increases considerably, though less than it does in girls. In some boys (between a fifth and a third of most groups studied) there is a distinct enlargement of the breast (sometimes unilaterally) about midway through adolescence. This usually regresses again after about one year.

First signs of puberty in the girl

In girls the start of breast enlargement—the appearance of the "breast bud"—is as a rule the first sign of puberty, though the appearance of pubic hair precedes it in about one-third. The uterus and vagina develop simultaneously with the breast. The labia and clitoris also enlarge. Menarche, the first menstrual period, is a late event in the sequence. Though it marks a definitive and probably mature stage of uterine development, it does not usually signify the attainment of full reproductive function. The early cycles may be more irregular than later ones and in some girls, but by no means all, are accompanied by discomfort. They are often anovulatory; that is, without the shedding of an egg. Thus there is frequently a period of adolescent sterility lasting a year to 18 months after menarche, but it cannot be relied on in the individual case. Similar considerations may apply to the male, but there is no reliable information about this. On average, girls grow about six centimetres (about 2.4 inches) more after menarche, though gains of up to twice this amount may occur. The gain is practically independent of whether menarche occurs early or late.

NORMAL VARIATIONS

Children vary a great deal both in the rapidity with which they pass through the various stages of puberty and in

Stages of puberty

the closeness with which the various events are linked together. At one extreme one may find a perfectly healthy girl who has not yet menstruated though her breasts and pubic hair are characteristic of the adult and she is already two years past her peak height velocity; and at the other, a girl who has passed all the stages of puberty within the space of two years.

In girls the interval from the first indication of puberty to complete maturity varies from 18 months to six years. The period from the moment when the breast bud first appears to menarche averages 2½ years, but it may be as little as six months or as much as 5½ years. The rapidity with which a child passes through puberty seems to be independent of whether puberty is occurring early or late. Menarche invariably occurs after peak height velocity has been passed.

In boys a similar variability of maturation occurs. The male genitalia may take between two and five years to attain full development, and some boys complete the whole process before others have moved from the first to the second stage.

The height spurt occurs relatively later in boys than in girls. Thus there is a difference between the average boy and girl of two years in age of peak height velocity but of only one year in the first appearance of pubic hair. Indeed, in some girls the acceleration in height is the first sign of puberty; this is never so in boys. A small boy whose genitalia are just beginning to develop can be unequivocally reassured that an acceleration in height is soon to take place, but a girl in the corresponding situation may already have had her height spurt.

Sex dimorphism. The differential effects on the growth of bone, muscle, and fat at puberty increase considerably the difference in body composition between the sexes. Boys have a greater increase not only in stature but especially in breadth of shoulders; girls have a greater relative increase in width of hips. These differences are produced chiefly by the changes that occur during puberty, but other sex differentiations arise before that time. Some, like the external genital difference itself, develop during fetal life. Others develop continuously throughout the whole growth period by a sustained differential growth rate. An example of this is the greater relative length and breadth of the forearm in the male when compared with whole arm length or whole body length.

Part of the sex difference in pelvic shape antedates puberty. Girls at birth already have a wider pelvic outlet. Thus the adaptation for childbearing is present from an early age. The changes at puberty are concerned more with widening the pelvic inlet and broadening the much more noticeable hips.

Physical and behavioral interaction. Children vary greatly in their tempo of growth. The effects are most dramatically seen at adolescence, but they are present at all ages from birth and even before.

Developmental and chronological ages

The concept of developmental age, as opposed to chronological age, is an important one. To measure developmental age, there is need of some way of determining how far along his own path to maturity a given child has gone. Therefore, there is need of a measure in which everyone at maturity ends up the same (not different as in height). The usual measure used is skeletal maturity or bone age. This is measured by taking an X-ray of the hand and wrist. The appearances of the developing bones can be rated and formed into a scale of development; the scale is applicable to boys and girls of all genetic backgrounds, though girls on average reach any given score at a younger age than do boys; and blacks on average, at least in the first few years after birth, reach a given score younger than do whites. Other areas of the body may be used if required. Skeletal maturity is closely related to the age at which adolescence occurs; that is, to maturity measured by some sex character developments. Thus the range of the chronological age within which menarche may normally fall is about 10 to 16½, but the corresponding range of bone age for menarche is only 12 to 14½. Evidently the physiological processes controlling progression of skeletal development are in most instances closely linked with those that initiate the events of adolescence. Furthermore, children tend to be consistently advanced or retarded during their whole growth period, at any rate after about age three.

There is little doubt that being an early or a late maturer may have repercussions on behaviour and that in some children these repercussions may be considerable. There is little enough solid information on the relation between emotional and physiological development, but what there is supports the common-sense notion that emotional attitudes are clearly related to physiological events.

Effects of late and early maturity

Larger size and earlier maturation. The rate of maturing and the age of onset of puberty are dependent on a complex interaction of genetic and environmental factors. Where the environment is good, most of the variability in age at menarche in a population is due to genetical differences. In many societies puberty occurs later in the poorly off, and, in most societies investigated, children with many siblings grow more slowly than children with few.

Modern tendency toward earlier maturity

During the last hundred years there has been a striking tendency for children to become progressively larger at all ages. This is known as the "secular trend." The magnitude of the trend in Europe and America is such that it dwarfs the differences between socioeconomic classes.

The data from Europe and America agree well: from about 1900, or a little earlier, to the present, children in average economic circumstances have increased in height at age five to seven by about one to two centimetres (0.4 to 0.8 inch) per decade, and at 10 to 14 by two to three centimetres (0.8 to 1.2 inches) each decade. Preschool data show that the trend starts directly after birth and may, indeed, be relatively greater from age two to five than subsequently. The trend started, at least in Britain, as early as 1850.

Most of the trend toward greater size in children reflects a more rapid maturation; only a minor part reflects a greater ultimate size. The trend toward earlier maturing is best shown in the statistics on age at menarche. A selection of the best data is illustrated in Figure 38. The trend is between three and four months per decade since 1850 in average sections of western European populations. Well-off persons show a trend of about half of this magnitude, having never been so retarded in menarche as the worse

From J. Tanner, *Growth at Adolescence* (1962); Blackwell Scientific Publications

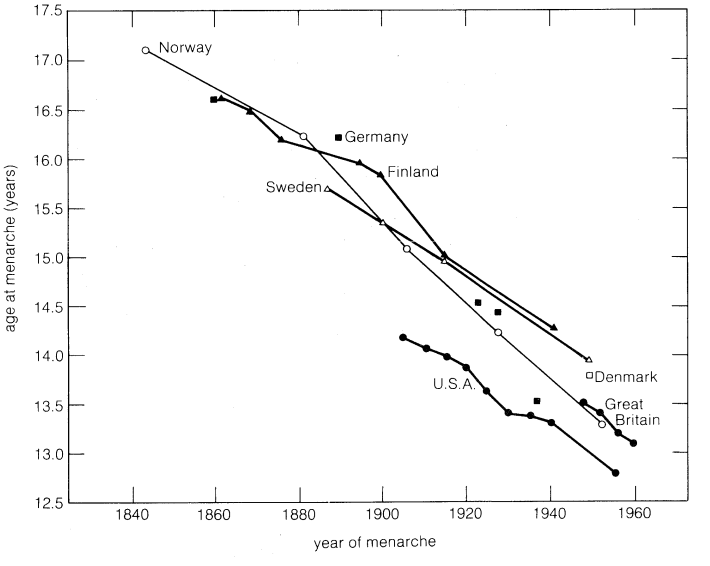


Figure 38: Secular trend in age at menarche, 1830–1960.

off. The causes of the secular trend are probably multiple. Certainly better nutrition is a major one and perhaps in particular more protein and calories in early infancy. A lessening of disease may also have contributed. Hot climates used to be cited as a potent cause of early menarche, but it seems that their effect, if any, is considerably less than that of nutrition. Some authors have supposed that the increased psychosexual stimulation consequent on modern urban living has contributed, but there is no positive evidence for this.

Hormones and growth

The main hormones concerned with growth are pituitary growth hormone, thyroid hormone, the sex hormones testosterone and estrogen, and the pituitary gonadotropic (sex-gland-stimulating) hormones.

Action of growth hormone

Pituitary growth hormone, a protein with molecular weight of 21,600 and of known amino-acid composition, is secreted by the pituitary gland throughout life. Exactly what its function is in the adult is not clear, but in the child it is necessary for growth; without it dwarfism results. During fetal life it seems not to be necessary, though normally present. It is not secreted at a constant rate all day but in small bursts of activity. Secretion by the pituitary is controlled by a substance sent to it from an adjacent part of the brain. The normal stimulus for secretion is not certain, but a sharp and "unnatural" lowering of blood sugar will cause growth hormone to be secreted, and this is used as a test. The hormone decreases the amount of fat and causes protein to be laid down in muscles and viscera. Children who lack it are fat as well as small; when given it by injection, they lose fat and grow rapidly.

The hormone is peculiar in being species-specific; that is, only growth hormone from human glands is active in man. Supplies of the hormone for treating children who need it are obtained at autopsy, and supply has been limited by this. Recombinant DNA technology shows possibilities in increased manufacture of this hormone in the laboratory.

Thyroid hormone from the thyroid gland in the neck is necessary for normal growth, though it does not itself stimulate growth, for example, in the absence of pituitary growth hormone. Without thyroid hormone, however, cells do not develop and function properly, especially in the brain. Babies who lack thyroid hormone at birth are small and have insufficiently developed brains; they are known as cretins. Frequently, if the condition is diagnosed and they are treated with thyroid hormone at once, they recover completely; the longer they go without treatment, the more likely it is that the brain damage will be permanent.

Thyroid lack may also develop later in childhood, when it causes a slowing of growth rate; full catch-up follows prompt treatment.

Testosterone, secreted by the interstitial cells of the testis, is important not only at puberty but before. Its secretion by the fetal testis cells is responsible for the development of certain parts of the male genital apparatus. If testosterone is not secreted at a particular and circumscribed time, the genitalia develop into the female form.

Effects of sex hormones on growth

Only small amounts of testosterone circulate between birth and puberty, but at puberty the interstitial cells develop greatly in response to pituitary luteinizing hormone (see below), and testosterone is secreted in large amounts, bringing about most of the changes of male puberty. It acts on a widespread series of receptors—for example, the cells of the penis, the muscles, the skin of the face, the cartilages of the shoulder, and certain parts of the brain. Most of the adolescent growth spurt is due to testosterone.

The female sex hormones, collectively called estrogens, are first secreted in quantity at puberty by cells in the ovary. They cause growth of the uterus, vagina, and breast; they act also on the bones of the hip, causing the specifically female widening. The adolescent growth spurt in the female is probably caused by testosterone-like substances (androgens) secreted by the adrenal gland in both male and female.

Initiation of puberty

The pituitary secretes two other hormones concerned in development: one, follicle-stimulating hormone (FSH), causes growth of the main portions of the ovary in the female and the sperm-producing cells in the testis of the male; the other, luteinizing hormone (LH), causes growth and secretion of the testosterone-secreting cells of the male and has an action in controlling the menstrual cycle in the female. The pituitary is caused to secrete gonadotropins by substances called releasing factors that come to it from adjacent areas of the brain, where they are made. Certain children develop all the changes of puberty, up to and including sperm production or ovulation, at an early age, either as the result of a brain lesion or as an isolated developmental, sometimes genetic, defect. The youngest mother on record was such a child; she gave birth to a full-term healthy infant by cesarean section at the age of five years and eight months. The existence of precocious puberty and the results of accidental ingestion by small children of male or female sex hormones indicate that breasts, uterus, and penis will respond to hormonal stimulation long before puberty. Evidently an increased end-organ sensitivity plays at most a minor part in puberal events.

The role of the hypothalamus in puberty

The signal to start the sequence of events is given by the brain, not the pituitary. Just as the brain holds the information on sex, so it holds information on maturity. The pituitary gland of a newborn rat successfully grafted in place of an adult pituitary begins at once to function in an adult fashion and does not have to wait until its normal age of maturation has been reached. It is the hypothalamus in the brain, not the pituitary, that must mature before puberty begins. Small amounts of sex hormones circulate from the time of birth, and these appear to inhibit the prepuberal hypothalamus from producing gonadotropin releasers. At puberty the hypothalamic cells become less sensitive to sex hormones. The small amount of sex hormones circulating then fails to inhibit the hypothalamus; gonadotropins are released, and these stimulate the production of testosterone by the testis or estrogen by the ovary. The level of the sex hormone rises until the same feedback circuit is re-established but now at a higher level of gonadotropins and sex hormones. The sex hormones are now high enough to stimulate the growth of secondary sex characters and to support mating behaviour.

Numerous factors may retard maturation or prevent normal growth, including hormonal disorders, metabolic defects, hereditary conditions, and inadequate nutrition. For information about specific growth disorders, see ENDOCRINE SYSTEMS: *Diseases and disorders of the human endocrine system*; METABOLISM; NUTRITION: *Nutritional diseases and disorder.* (J.M.T./Ed.)

AGING AND SENESCENCE

Life-span

Definition of life-span

It is a commonplace that organisms die, some after only a brief existence, like that of the mayfly whose adult life burns out in a day, and others like that of the gnarled bristlecone pines that have lived thousands of years. The limits of the life-span of each species appear to be determined ultimately by heredity. Locked within the code of the genetic material are instructions that specify the age beyond which a species cannot live given even the most favourable conditions. And many environmental factors act to diminish that upper age limit.

MEASUREMENT OF LIFE-SPAN

The maximum life-span is a theoretical number whose exact value cannot be determined from existing knowledge about an organism; it is often given as a rough estimate based on the longest lived organism of its species known to date. A more meaningful measure is the average life-span; this is a statistical concept that is derived by the analysis of mortality data for populations of each species. A related term is the expectation of life, a hypothetical number computed for humans from mortality tables drawn up by insurance companies. Life expectancy represents the average number of years that a group of persons, all born at the same time, might be expected to live, and it is based on the changing death rate over many past years.

The concept of life-span implies that there is an individual whose existence has a definite beginning and end. What constitutes the individual in most cases presents no problem: among organisms that reproduce sexually the in-

dividual is a certain amount of living substance capable of maintaining itself alive and endowed with hereditary features that are in some measure unique. In some organisms, however, extensive and apparently indefinite growth takes place and reproduction may occur by division of a single parent organism, as in many protists, including bacteria, algae, and protozoans. If these divisions are incomplete, a colony results; if the parts separate, genetically identical organisms are formed. In order to consider life-span in such organisms, the individual must be defined arbitrarily since the organisms are continually dividing. In a strict sense, the life-spans in such instances are not comparable to those forms that are sexually produced.

The beginning of an organism can be defined by the formation of the fertilized egg in sexual forms; or by the physical separation of the new organism in asexual forms (many invertebrate animals and many plants). In animals generally, birth is considered to be the beginning of the life-span. The timing of birth, however, is so different in various animals that it is only a poor criterion. In many marine invertebrates the hatchling larva consists of relatively few cells, not nearly so far along toward adulthood as a newborn mammal. For even among mammals, variations are considerable. A kangaroo at birth is about an inch long and must develop further in the pouch, hardly comparable to a newborn deer, who within minutes is walking about. If life-spans of different kinds of organisms are to be compared, it is essential that these variations be accounted for. The end of an organism's existence results when irreversible changes have occurred to such an extent that the individual no longer actively retains its organization. There is thus a brief period during which it is impossible to say whether the organism is still alive, but this time is so short relative to the total length of life that it creates no great problem in determining life-span.

Potential immortality

Some organisms seem to be potentially immortal. Unless an accident puts an end to life, they appear to be fully capable of surviving indefinitely. This faculty has been attributed to certain fishes and reptiles, which appear to be capable of unlimited growth. Without examining the various causes of death in detail (see DEATH) a distinction can be made between death as a result of internal changes (*i.e.*, aging) and death as a result of some purely external factor, such as an accident. It is notable that the absence of aging processes is correlated with the absence of individuality. In other words, organisms in which the individual is difficult to define, as in colonial forms, appear not to age. (P.W.F.)

Table 3: Maximum Ages for Some Seed Plants

plant	maximum age in years		locale of verified specimen
	estimated	verified	
Conifers			
Common juniper (*Juniperus communis*)	2,000	544	Kola Peninsula, northeastern Russia
Norway spruce (*Picea abies*)	1,200	350–400	Eichstätt, Bavaria
European larch (*Larix decidua*)	700	417	Riffel Alp, Switz.
Scotch pine (*Pinus sylvestris*)		584	
Swiss stone pine (*Pinus cembra*)	1,200	750	Riffel Alp, Switz.
White pine (*Pinus strobus*)		400–450	
Bristlecone pine (*Pinus aristata*)		4,900	Wheeler Peak, Humboldt National Forest, Nevada
Sierra redwood (*Sequoiadendron giganteum*)	4,000	2,200–2,300	Northern California
Flowering plants			
Monocots			
Dragon tree (*Dracaena draco*)	200*		Tenerife, one of the Canary Islands
Solomon's-seal (*Polygonatum*)	17†		
Dicots			
Dwarf birch (*Betula nana*)		80	Eastern Greenland
European beech (*Fagus sylvatica*)	900	250	Montigny, Normandy, France
English oak (*Quercus robur*)	2,000	1,500	Hasbruch Forest, Lower Saxony
Bo tree (*Ficus religiosa*)	2,000–3,000‡		Buddh Gaya, India; Anuradhapura, Ceylon
Linden (*Tilia*)		815	Lithuania
English ivy (*Hedera helix*)	440		Ginac, near Montpellier, France

*Exaggerated estimates for this historic specimen reach 6,000 years. †Scars on root-stock counted. ‡According to Buddhist and Roman history.

PLANTS

Plants grow old as surely as do animals; however, a generally accepted definition of age in plants has not yet been realized. If the age of an individual plant is that time interval between the reproductive process that gave rise to the individual and the death of the individual, the age attained may be given readily for some kinds of plants but not for others. Table 3 lists maximum ages, both estimated and verified, for some seed plants.

Problem of defining age. An English oak that has 1,000 annual rings in the trunk is 1,000 years old. But age is less certain in the case of an arctic lupine that germinated from a seed that, containing the embryo, had been lying in a lemming's burrow in the arctic permafrost for 10,000 years.

The mushroom caps that appear overnight last for only a few days, but the network of fungus filaments in the soil (the mycelia) may be as old as 400 years. Because of important differences in structure, the life-span of higher plants cannot be compared with that of higher animals. Normally, embryonic cells (that is, cells capable of changing in form or becoming specialized) cease to exist very early in the life of an animal. In plants, however, embryonic tissue—the plant meristems—may contribute to growth and tissue formation for a much longer time, in some cases throughout the life of the plant. Thus the oldest known trees, bristlecone pines of California and Nevada, have one meristem (the cambium) that has been adding cells to the diameter of these trees for, in many cases, more than 4,000 years and another meristem (the apical) that has been adding cells to the length of these trees for the same period. These meristematic tissues are as old as the plant itself; they were formed in the embryo. The wood, bark, leaves and cones, however, live for only a few years. The wood of the trunk and roots, although dead, remains a part of the tree indefinitely, but the bark, leaves, and cones are continually in the process of dying and sloughing off.

Ageless embryonic tissue

Among the lower plants only a few mosses possess structures that enable an estimate of their age to be made. The haircap moss (*Polytrichum*) grows through its own stem tip each year, leaving a ring of scales that marks the annual growth. Three to five years' growth in this moss is common, but life-spans of 10 years have been recorded. The lower portions of such a moss are dead, though intact. Peat moss (*Sphagnum*) forms extensive growths that fill acid bogs with a peaty turf consisting of the dead lower portions of mosses whose living tops continue growing. Mosses that become encrusted with lime (calcium carbonate) and form "tufa" beds several metres thick also have living tips and dead lower portions. On the basis of their observed annual growth, some tufa mosses are estimated to have been growing for as long as 2,800 years.

No reliable method for determining the age of ferns exists, but on the basis of size attained and growth rate, some tree ferns are thought to be several decades old. Some club mosses, or lycopsids, have a "storied" growth pattern similar to that of the haircap moss. Under favourable conditions some specimens live five to seven years.

The woody seed plants, such as conifers and broadleaf trees, are the most amenable to determination of age. In temperate regions, where each year's growth is brought to an end by cold or dryness, every growth period is limited by an annual ring—a new layer of wood added to the diameter of the tree. These rings may be counted on the cut ends of a tree that has been felled or, using a special instrument, a cylinder of wood can be cut out and the growth rings counted and studied. In the far north growth rings are so close together that they are difficult to count.

In the moist tropics growth is more or less continuous, so that clearly defined rings are difficult to find.

Errors in estimating age

Often the age of a tree is estimated on the basis of its diameter, especially when the average annual increase in diameter is known. The source of greatest error in this method is the not infrequent fusing of the trunks of more than one tree, as, for example, occurred in a Montezuma cypress in Santa María del Tule, a little Mexican village near Oaxaca. This tree, described by the Spanish explorer Hernan Cortés in the early 1500s, was earlier estimated on the basis of its great thickness to be 6,000 years old; later studies, however, proved it to be three trees grown together. Estimates of the age of some English yews have been as high as 3,000 years, but these figures, too, have turned out to be based on the fusion of close-growing trunks, none of which is more than 250 years old. Increment borings of bristlecone pines have shown specimens in the western United States to be 4,600 years old.

Growing season of seed plants. *Annuals.* Plants, usually herbaceous, that live for only one growing season and produce flowers and seeds in that time are called annuals. They may be represented by such plants as corn and marigolds, which spend a period of a few weeks to a few months rapidly accumulating food materials. As a result of hormonal changes—brought about in many plants by changes in environmental factors such as day length and temperature—leaf-producing tissues change abruptly to flower-producing ones. The formation of flowers, fruits, and seeds rapidly depletes food reserves and the vegetative portion of the plant usually dies. Although the exhaustion of food reserves often accompanies death of the plant, it is not necessarily the cause of death.

Biennials. These plants, too, are usually herbaceous. They live for two growing seasons. During the first season, food is accumulated, usually in a thickened root (beets, carrots); flowering occurs in the second season. As in annuals, flowering exhausts the food reserves, and the plants die after the seeds mature.

Perennials. These plants have a life-span of several to many years. Some are herbaceous (iris, delphinium), others are shrubs or trees. The perennials differ from the above-mentioned groups in that the storage structures are either permanent or are renewed each year. Perennials require from one to many years growth before flowering. The preflowering (juvenile) period is usually shorter in trees and shrubs with shorter life-spans than in those with longer life-spans. The long-lived beech tree (*Fagus sylvatica*), for example, passes 30–40 years in the juvenile stage, during which time there is rapid growth but no flowering.

Some plants—cotton and tomatoes, for example—are perennials in their native tropical regions but are capable of blooming and producing fruits, seeds, or other useful parts in their first year. Such plants are often grown as annuals in the temperate zones.

Longevity of seeds. Although there is great variety in the longevity of seeds, the dormant embryo plant contained within the seed will lose its viability (ability to grow) if germination fails to occur within a certain time. Reports of the sprouting of wheat taken from Egyptian tombs are unfounded, but some seeds do retain their viability a long time. Indian lotus seeds (actually fruits) have the longest known retention of viability. On the other hand, seeds of some willows lose their ability to germinate within a week after they have reached maturity.

The oldest living seeds

The loss of viability of seeds in storage, although hastened or retarded by environmental factors, is the result of changes that take place within the seed itself. The changes that have been investigated are: exhaustion of food supply; gradual denaturing or loss of vital structure by protoplasmic proteins; breakdown of enzymes; accumulation of toxins resulting from the metabolism of the seed. Some self-produced toxins may cause mutations that hamper seed germination. Since seeds of different species vary greatly in structure, physiology, and life history, no single set of age factors can apply to all seeds. (L.K.)

ANIMALS

Much of what is known of the length of life of animals other than man derives from observations of domesticated species in laboratories and zoos. One has only to consider how few animals reveal their age to appreciate the difficulties involved in answering the apparently simple question of how long they live in nature. In many fishes, a few kinds of clams, and an occasional species of other groups, growth is seasonal, so that annual zones of growth, much like tree rings, are produced in some part of the organism. Among game species, methods of determining relative age by indicators such as the amount of tooth wear or changes in bone structure have yielded valuable information. Bird bands and other identifying marks also make age estimation possible. But one of the consequences of the fact that animals move is that very little is known about the life-span of most species as they exist in nature.

Maximum and average longevity. The extreme claims of longevity that are occasionally made for one species or another have consistently been proven false when subjected to critical scrutiny. Although the maximum life-span that has been observed for a particular species cannot be considered absolute, since a limited number of individuals at best has been studied, this datum probably provides a fair approximation of the greatest age attainable for this kind of animal under favourable conditions. Animals in captivity, which provide most of the records of extreme age, are exposed to far fewer hazards than those in the wild. In the accompanying table of maximum longevity, particular species have been so selected as to encompass the known range of longevity of other members of the taxonomic group to which they belong.

Table 4: Maximum Longevity of Animals in Captivity*

animal	life-span in years	animal	life-span in years
Mammals		**Amphibians**	
Bat (*Eptesicus fuscus*)	2	European black salamander (*Salamandra atra*)	3
Grizzly bear (*Ursus horribilis*)	31	Spotted salamander (*Ambystoma maculatum*)	25
Cat (*Felis catus*)	21	Frog (*Rana* species)	5–15
Chimpanzee (*Pan troglodytes*)	37	**Fishes**	
Dog (*Canis familiaris*)	34	Eel (*Anguilla rostrata*)	6
Elephant, Indian (*Elephas maximus*)	57	Goldfish (*Carassius auratus*)	25
Goat (*Capra hircus*)	18	Sturgeon (*Acipenser transmontanus*)	50
Golden hamster (*Mesocricetus auratus*)	1.8	**Insects**	
Horse (*Equus caballus*)	62	Ant (*Lasius* species)	15
Lion (*Panthera leo*)	29	Buprestid beetle (*Buprestis splendens*)	30
Mouse (*Mus musculus*)	3	Fruit fly (*Drosophila melanogaster*)	0.1
Ox (*Bos taurus*)	30	**Arachnids**	
Squirrel, gray (*Sciurus carolinensis*)	15	Bird spider (*Avicularis avicularis*)	15
Wild boar (*Sus scrofa*)	27	Rocky Mountain wood tick (*Dermacentor andersoni*)	3-4
Birds		**Crustaceans**	
Blue jay (*Cyanocitta cristata*)	4	Crayfish (*Astacus fluviatilis*)	30
Canary (*Serinus canaria*)	24	Water flea (*Daphnia magna*)	0.2
Macaw (*Ara macao*)	64	**Mollusks**	
Nightingale (*Luscinia luscinia*)	3.8	Clams, various species	1–10
Pigeon (*Columba livia domestica*)	35	Snails, various species	1–30
Titmouse (*Parus major*)	9	**Annelid worms**	
Reptiles		Earthworm (*Lumbricus terrestris*)	10
Alligator (*Alligator mississipiensis*)	56	Medicinal leech (*Hirudo medicinalis*)	27
Garter snake (*Thamnophis sirtalis*)	6	**Rotifers**	
Box turtle (*Terrapene carolina*)	123	Various species	0.03–0.1
Giant tortoise (*Testudo elephantopus*)	177		
Water turtle (*Pseudemys scripta*)	7		

*Condensed and adapted from W.S. Spector (ed.), *Handbook of Biological Data*, 1956.

Environmental influences. Life-span usually is measured in units of time. Although this may seem eminently logical, certain difficulties may arise. In cold-blooded animals in general, the rate of metabolism that determines the various life processes varies with the temperatures to which they are exposed. If aging depends on the expenditure of a fixed amount of vital energy, an idea first proposed in 1908, life-span will vary tremendously depending on

Temperature effects

temperature or other external variables that influence life-span. There is considerable evidence attesting at least to the partial cogency of this argument. So long as a certain range is not exceeded, cold-blooded invertebrates do live longer at low than at high temperatures. Rats in the laboratory live longest on a somewhat restricted diet that does not permit maximum metabolic rate. Of perhaps even greater significance is the fact that many animals undergo dormant periods. Many small mammals hibernate; a number of arthropods have life cycles that include periods during which development is arrested. Under both conditions the metabolic rate becomes very low. It is questionable whether such periods should be included in computing the life-span of a particular organism. Comparisons between species, some of which have such inactive periods while others do not, are dangerous. It is possible that life-span could be measured more adequately by total metabolism; however, the data that are necessary for this purpose are almost entirely lacking.

Length of life is controlled by a multitude of factors, which collectively may be termed environment, operating on a genetic system that determines how the individual will respond. It is impossible to list all the environmental factors that may lead to death. For analytical purposes it is, however, useful to make certain formal separations. Every animal is exposed to (1) a pattern of numerous events, each with a certain probability of killing the individual at any moment and, in the aggregate, causing a total probability of death or survival; (2) climatic and other changes in the habitat, modifying the frequency with which the various potentially fatal events occur; and (3) progressive systemic change, inasmuch as growth, reproduction, development, and senescence are characteristics intrinsic in the organism and capable of modifying the effects of various environmental factors.

Patterns of survival. Consider a group of similar animals of the same age. Although no two individuals can have precisely the same environment, let it be assumed that the environment of the group remains effectively constant. If the animals undergo no progressive physiological changes, the factors causing death will produce a death rate that will remain constant in time. Under these conditions, it will take the same amount of time for the population to become reduced to one-half its former number, no matter how many animals remain at the beginning of the period considered. The animals therefore survive according to the pattern of an accident curve. This is the sense in which many of the lower animals are immortal. Although they die, they do not age; how long they have already lived has no influence on their further life expectation.

Another group of animals may consist of individuals that differ markedly in their responses to the constant environment. They may be genetically different, or their previous development may have caused variations to arise. Those individuals that are most poorly suited to the new environment will die, leaving survivors that are better adapted. The same result can also be achieved in other ways. If the environment varies geographically, those individuals that happen to find areas in which existence can be maintained will survive, while the remainder will die. Or, as a result of their own properties, animals in a constant environment may acclimate in a variety of ways, thus adjusting to the existing conditions. The pattern of survival that results in each of these cases is one in which the death rate declines with time, as illustrated by the selection–acclimation curve.

In the absence of death from other causes, all members of a population may exist in their environment until the onset of senescence, which will cause a decline in the ability of individuals to survive. In a sense they can be considered to wear out as does a machine. Their survival is best described by individual differences among members of the population that determine the curvature of the survival line (wearing-out curve). The more the population varies, the less abrupt is the transition from total survival to total death.

Under the actual conditions of existence of animals the three types of survival (accident pattern, selection–acclimation pattern, wearing-out pattern) above all enter as components of the realized survival pattern. Thus in animals that are carefully maintained in the laboratory, survival is approximately that of the wearing-out pattern. Environmental accidents can be kept to a minimum under these conditions, and survival is almost complete during the major part of the life-span. In all known cases, however, the early stages of the life-span are characterized by a noticeable contribution of the selection–acclimation pattern. This must be interpreted as a result of developmental changes that accompany the early life of the individuals and of selective processes that operate on those organisms whose genetic constitutions are ill fitted for that environment.

In some of the larger mammals in nature, the existing evidence points to a similar survival pattern. In a variety of other animals, however, and including fishes and invertebrates, mortality in the young stages is so high that the selection–acclimation curve predominates. One estimate places the mortality of the Atlantic mackerel during its first 90 days of life as high as 99.9996 percent. Since some mackerel do live for several years, a mortality rate that decreases with age is indicated. Similar considerations probably apply to all those animals that have larval stages that serve as dispersal mechanisms.

Highest mortality in aquatic animals

When the postjuvenile portion of the life-span is considered by itself, a number of animals for which such information has been gathered—including primarily fishes and birds—have survivorship curves that are dominated by the accident pattern. In these species in nature, death from old age apparently is rare. Their chance of surviving to an advanced age is so small that it may be statistically negligible. In modern times, human predation is a large factor in the mortality of these species in many cases. Since deaths from fishing and hunting are largely independent of age, once an animal has reached a certain minimum size, such a factor only makes the survival curve steeper but does not change its shape. One consequence of such increased mortality is that fewer old and large individuals are noticed in a population.

More complex survival patterns, such as the hypothetical one illustrated, undoubtedly exist. They should be looked for in those species in which extensive reorganization of the animal is part of the normal life cycle. In effect, these animals change their environment radically, in some cases several times during a lifetime. The frog offers a familiar example. During its period of early development and until shortly after hatching, the animal is subject to major internal, and some external, change. As a tadpole it is adjusted to an aquatic, herbivorous life. The metamorphosis to the terrestrial, carnivorous adult form is accompanied by varied physiological stresses that must be expected to produce a temporary increase in mortality rate. In some insects the eggs, larvae, pupae, and adults are exposed to and respond to quite different environments, and a survivorship pattern even more complex than that described by the composite curve may exist.

Effects of metamorphosis

The same species will exhibit changed survival in different environments. In captivity an animal population may approach the wearing-out pattern; in its natural habitat survivorship may vary with age in a quite different way. Although one can assign a maximum potential life-span to an individual—while realizing that this maximum may not be attained—it is impossible to specify the survivorship pattern unless the environment is also specified. This is another way of saying that life-span is the joint property of the animal and the environment in which it lives.

HUMAN LIFE-SPAN

The exact duration of human life is unknown, although there is presumably a maximum life-span for the human race established in the genetic material. At first thought, this statement seems irrational. Surely no human being can live 1,000 years. Even though all may agree that the likelihood of an individual living 1,000 years is infinitesimal, there is no scientific proof that this statement is or is not true. The indeterminacy of the maximum limit of human life is made more comprehensible if one chooses a number that may appear to be a more reasonable limit.

Since there is no verified instance of a person having

lived 150 years, this number may, for purposes of illustration, be arbitrarily accepted as the maximum limit of the span of human life. But if the possibility is admitted that an individual may live exactly 150 years, there is no valid reason for rejecting the possibility that some other individual may live 150 years and one minute. And if 150 years and one minute is accepted, why not 150 years and two minutes, and so on? Thus, based on existing knowledge of longevity, a precise figure for the span of human life cannot be given.

Studies on longevity. Much information concerning the inheritance of longevity has come from the study of genealogical records of nobility and landed gentry. The early genealogical studies were criticized on the grounds that the downward trend in the death rate (attributable generally to scientific advancements) introduced a spurious correlation in statistics derived from records extending over long periods of time. It was argued that in some instances records were included of persons who, at the time of the study, had not had the opportunity of living out their possible life-span. The general finding of such investigations was that the expectation of life of sons of long-lived parents (*i.e.,* those living to age 70 years or older) was greater than that of sons of shorter-lived parents (*i.e.,* those having attained less than age 50 at the time of death).

An American biostatistician attempted to avoid the defects of genealogical studies by collecting records of the family histories of 365 nonagenarians (90-year-old persons) and of a comparison group of 143 individuals of varying ages, selected because all of their six immediate ancestors were dead. The study introduced the concept of "total immediate ancestral longevity," or TIAL—the sum of the ages at death of the two parents and the four grandparents of a given person—as a measure of longevity. This number is unlikely to be greater than 600 or less than 90. The average TIAL of the nonagenarians and centenarians definitely exceeded that of the comparison group. This held true not only for the six immediate ancestors as a group but also for each category—father, mother, paternal and maternal grandparents. In the same study, investigators also computed the expectation of life for sons of fathers as classified in three groups by age at death: (1) under age 50, (2) from age 50 to age 79, and (3) age 80 or over. The expectation of life for the three groups at birth was 47.0, 50.5, and 57.2 years, respectively. The same relative ranking continued through the lifetime of the sons, their expectation of life at age 40 being 27.3, 28.9, and 32.0 years, respectively.

The TIAL concept

While certain doubts have been raised about the validity of these as well as earlier studies, taken at their face value, these data show clearly that long-lived persons had parents and grandparents who lived longer than the parents and grandparents of shorter lived persons.

Since longevity is important in life insurance underwriting, several studies have been made of the relationship between heredity and the life-span by an analysis of life insurance records. Such analyses showed that policyholders both of whose parents were living when the policy was written live longer than those whose parents were dead when the policy was written. These results are in conformity with those obtained from genealogical records and family histories.

Influence of heredity

Each of the various types of studies of the inheritance of longevity—genealogical records, life insurance records, and family histories of the general population—has limitations that restrict the applicability of the findings. The principal studies indicate, nevertheless, that the children of long-lived parents are more likely to be long-lived than are the children of short-lived parents. Conversely, the immediate ancestors—parents and grandparents—of long-lived persons on the average are older at death than are the immediate ancestors of persons who die at a relatively young age. These studies support the conclusion, mentioned earlier, that longevity is determined in part by heredity.

Actual versus possible life-span. It should be observed that this conclusion relates to the inheritance of longevity—the observed expression of the span of life—and not to the span of life itself. The actual length of life itself is shorter than the possible life-span, since the former reflects the effect of unfavourable environmental factors. In the absence of any biological data from which the maximum limit of the span of life can be determined precisely, an estimate of the limit must be obtained from observation of the actual length of life of persons who already have died. But such observations cannot establish a fixed limit for the span of life.

The estimation of the length of the span of life from observed data is a form of sampling from a large but incomplete population. The tabulation of the ages at death of a large number of persons from a large general population of the United States will give an asymmetrical frequency distribution with two modes, or peaks, of highest frequency: the first at age less than one year and the second between ages 75 and 80 years. The frequency distribution is bounded by age zero at the lower limit but there is no boundary at the upper limit. The number of deaths of persons whose length of life is near the upper limit of this frequency distribution (*e.g.,* 100 years or more) varies from year to year. The age of the oldest person dying also varies from year to year.

The number of deaths of centenarians (100-year-old persons) depends in part upon the number of deaths counted. Ages at death are frequently unverified, so that the true numbers of centenarians almost certainly deviate from those given in official vital statistics. Moreover, only a very small proportion of the deaths that have occurred throughout the history of the human race have been registered. The potential number of future deaths greatly exceeds the number that already has occurred. Statistical theory supports the expectation that as the total number of deaths continues to increase, the death of a person whose length of life will be longer than that of any person previously known will be recorded.

Observation of the length of life of persons who have died can show that it is possible for a human being to live to the oldest age recorded as of any specified date and can provide an estimate of the relative frequency or probability of that event. But such observations do not provide a logical basis for fixing any age as the maximum possible limit of the life-span.

The continuation of the worldwide decline in the death rate will naturally result in an increase in the number of persons who live until age 100 years or more. Since the number of persons who may live to an advanced age, such as 110 or 115 years, is directly related to the number of persons who live to age 100, an increase in the latter number will increase the probability that the death of an individual attaining some greater age (*e.g.,* 115 years) will be recorded at some future date.

Many instances of persons alleged to have died at an age considerably greater than 100 years have been recorded. Statements concerning the age at death of biblical characters such as Methuselah can be dismissed, since scientific verification is impossible. Three of the most frequently cited cases of more recent times are: Thomas Parr, who died in November 1635 at the alleged age of 152 years; Henry Jenkins, who died in December 1670 at the alleged age of 169 years; and Catherine, countess of Desmond, who died in 1604 at the alleged age of 140 years. William Harvey, a famous English physician, performed an autopsy on Thomas Parr and the account of the autopsy was cited for many years as evidence that Harvey—in his paper—had confirmed Parr's age. Quite apart from the fact that it is impossible accurately to determine the age of a person by an autopsy, Harvey made no attempt to verify Parr's age but merely referred to the current estimates. Subsequent investigations have revealed that no proof exists of the age at death of any of these three individuals and that their reported ages were based solely upon hearsay.

Famous age records

An example with more definite documentation is that of Christian Jacobsen Drakenberg, stated to have been born on November 18, 1626, and to have died on October 9, 1772, aged 145 years and 325 days. Although the authenticity of his age was attested to by many persons, including two celebrated Scandinavian actuaries, later investigations cast doubt upon the record. It is difficult to accept the statements concerning Drakenberg's age at death, since

this age is more than 30 years greater than the next oldest verified age at death—a difference that in itself casts doubt on its authenticity.

Of eight individuals for whom records substantiate the fact that each had lived more than 108 years, seven were females. Six of the eight were more than 110 years old at death. The oldest was Pierre Joubert, who was born July 15, 1701, and died November 16, 1814, aged 113 years and 124 days. Discounting the Drakenberg record, this is the oldest age at death that has been generally accepted as authentic.

It may be concluded that the span of human life is at least 114 years, but that this is not the maximum upper limit. This does not mean the span of life of each individual now living or to be born in the future is at least 114 years. The span of life, since it is determined by heredity, varies from one individual to another as do other genetically determined traits.

Factors influencing life-span of man

A significant proportion of human embryos and fetuses die before birth. Other infants at birth have defects that limit their span of life to a few years. Some malformations (*e.g.,* certain cardiovascular defects) are developmental rather than genetic in the strict sense of the word and can be corrected so that the length of life of such persons is extended.

In the past the length of life of most individuals has been considerably shorter than their possible span of life because of unhealthful environmental factors. As these factors are increasingly brought under control or eliminated, the actual length of life will approach more closely the span of life. At the end of the 18th century the expectation of life at birth in North America and northwestern Europe was about 35 or 40 years. By 1970 it exceeded 70 years, and at some future date the death of a person at an authenticated age of more than 114 years can be expected.

There is no evidence that the span of human life has increased since the beginning of recorded history. Neither is there any evidence that the death rate of centenarians has decreased. The expected increase in the number of centenarians results from a decrease in the death rate at ages under 100 years and not from any demonstrable increase in the maximum length of the span of life. The remarkable increase in the average length of life during the past 2,000 years—from 20–25 years to 70 years under favourable conditions—has increased the likelihood that a person may live to the maximum limit of his span of life. (P.W.F.)

Aging: general considerations

Aging consists of the progressive changes that take place in a cell, an organ, or the total organism with the passage of time. It is a process that goes on over the entire adult life-span of any living thing. Gerontology, the study of the aging process, is devoted to the understanding and control of all factors contributing to the finitude of individual life. It is not concerned exclusively with debility, which looms so large in human experience, but deals with a much wider range of phenomena. Every species has a life history in which the individual life-span has an appropriate relationship to the reproductive life-span and to the mechanism of reproduction and the course of development. How these relationships evolved is as germane to gerontology as it is to evolutionary biology. It is also important to distinguish between the purely physicochemical processes of aging and the accidental organismic processes of disease and injury that lead to death.

Gerontology, therefore, can be defined as the science of the finitude of life as expressed in the three aspects of longevity, aging, and death, examined in both evolutionary and individual (ontogenetic) perspective. Longevity is the span of life of an organism. Aging is the sequential or progressive change in an organism that leads to an increased risk of debility, disease, and death; senescence consists of these manifestations of the aging process.

Actuarial functions

The viability (survival ability) of a population is characterized in two actuarial functions: the survivorship curve (A in Figure 39) and the age-specific death rate, or Gompertz function (B in Figure 39). The relation of such factors as aging characteristics, constitutional vigour, physical factors, diet, and exposure to disease-causing organisms to the actuarial functions is complex; there is, nevertheless, no substitute for them as measures of the aging process and of the effect of environmental or genetic modifiers.

The age-specific mortality rate is the most informative actuarial function for investigations of the aging process. It was first pointed out by an English actuary, Benjamin Gompertz, in 1825 that the mortality rate increases in geometric progression—*i.e.,* by a constant ratio in successive equal age intervals. Hence, a straight line, known as the Gompertz function, results when death rates are plotted on a logarithmic (ratio) scale. The prevalence of many diseases and disabilities rises in the same geometrical manner as does the mortality rate, important exceptions being some infectious diseases and diseases arising from disturbances of the immunological system. Although the life tables of most species are remarkably similar in form, even closely related species can differ markedly in the relative incidence of the major causes of death. (G.A.Sa.)

BIOLOGICAL THEORIES OF AGING

Aging has many facets. Hence there are a number of theories, each of which many explain one or more aspects of aging; there is, however, no single theory that explains all of the phenomena of aging.

Genetic theories. One theory of aging assumes that the life-span of a cell or organism is genetically determined—that the genes of an animal contain a "program" that determines its life-span just as eye colour is determined genetically. Although long life is recognized often as a familial characteristic, and short-lived strains of fruit flies, rats, and mice can be produced by selective breeding, other factors clearly can significantly alter the basic genetic program of aging.

Another genetic theory of aging assumes that cell death is the result of "errors" introduced in the formation of key proteins, such as enzymes. Slight differences induced in the transmission of information from the deoxyribonucleic acid (DNA) molecules of the chromosomes through ribonucleic (RNA) molecules (the "messenger" substance) to the proper assembly of the large and complex enzyme molecules could result in a molecule of the enzyme that would not "work" properly. These so-called error theories have not yet been firmly established, but studies are in progress.

Cell errors and mutations

As cells grow and divide, a small proportion of them undergo mutation; that is, they become "different" with a change in their chromosome structure that is then reproduced when they again divide. The "somatic mutation" theory of aging assumes that aging is due to the gradual accumulation of mutated cells that do not perform normally.

Non-genetic theories. Other theories of aging focus attention on factors that can influence the expression of a genetically determined "program." One of these is the "wear-and-tear" theory, which assumes that animals and cells, like machines, simply wear out. Animals, however, unlike machines, have some ability to repair themselves, so that this theory does not fit the facts of a biological system. A corollary to the wear-and-tear theory is the presumption that waste products accumulate within cells and interfere with function. The accumulation of highly insoluble particles, known as "age pigments," has been observed in muscle cells in the heart and nerve cells of both human beings and other animals.

With increasing age, tendons, skin, and even blood vessels lose elasticity. This is due to the formation of cross-links between or within the molecules of collagen (a fibrous protein) that give elasticity to these tissues. The "cross-linking" theory of aging assumes that similar cross-links form in other biologically important molecules, such as enzymes. These cross-links could alter the structure and shape of the enzyme molecules so that they are unable to carry out their functions in the cell.

Auto-immunity control theories

Another theory of aging assumes that immune reactions, normally directed against disease-producing organisms as well as foreign proteins or tissue, begin to attack cells of the individual's own body. In other words, the system that produces antibodies loses its ability to distinguish between

"self" and foreign proteins. This "autoimmune" theory of aging is based on clinical rather than on experimental evidence.

These theories all attempt to explain aging in terms of cellular and molecular changes. Actually, age changes are much more marked in the overall performance of an individual than in cellular processes that can be measured. The age decrement in the ability to perform muscular work is much greater than any changes that can be detected in the enzyme activities of the muscles that perform the work. It is possible that aging in an individual is actually due to a breakdown in the control mechanisms that are required in a complex performance. (N.W.S./Ed.)

NATURAL HISTORY OF AGING

Reproduction and aging. Reproduction is an all-important function of an organism's life history, and all other vital processes, including senescence and death, are shaped to serve it. The distinction between semelparous and iteroparous modes of reproduction is important for an understanding of biological aging. Semelparous organisms reproduce by a single reproductive act. Annual and biennial plants are semelparous, as are many insects and a few vertebrates, notably salmon and eels. Iteroparous organisms, on the other hand, reproduce recurrently over a reproductive span that usually covers a major part of the total life-span.

In semelparous forms, reproduction takes place near the end of the life-span, after which there ensues a rapid senescence that quickly leads to the death of the organism. In plants the senescent phase is usually an integral part of the reproductive process and essential for its completion. The dispersal of seeds, for example, is accomplished by processes—including ripening and fall (abscission) of fruits and drying of seed pods—that are inseparable from the overall senescence process. Moreover, the onset of plant senescence is invariably initiated by the changing levels of hormones, which are under systemic or environmental control. If, for example, the hormone auxin is prevented, by experimental means, from influencing the plant, the plant lives longer than normal and undergoes an atypical prolonged pattern of senescent change.

The significance of aging patterns in insects

Useful inferences can be drawn from the study of the aging processes of insects that display two distinct kinds of adaptive coloration: the procryptic, in which the patterns and colours afford the insect concealment in its native habitat; and the aposematic, in which the vivid markings serve as a warning that the insect is poisonous or bad tasting. The two adaptation patterns have different optimal species survival strategies: the procryptics die out as quickly as possible after completing reproduction, thus reducing the opportunity for predators to learn how to detect them; the aposematics have longer post-reproductive survival, thus increasing their opportunity to condition predators. Both adaptations are found in the family of saturniid moths, and it has been shown that the duration of their post-reproductive survival is governed by an enzyme system that controls the fraction of time spent in flight: procryptics fly more, exhaust themselves, and die quickly; aposematics fly less, conserve their energies, and live longer.

These examples indicate that in semelparous forms, in which full vigour and function are required until virtually the end of life, senescence has an onset closely coupled with the completion of the reproductive process and is governed by relatively simple enzymatic mechanisms that can be modified by natural selection. Such specific, genetically controlled senescence processes are instances of programmed life termination.

The iteroparous forms include most vertebrates, most of the longer-lived insects, crustaceans and spiders, cephalopod and gastropod mollusks, and perennial plants. In contrast to semelparous forms, iteroparous organisms need not survive to the end of their reproductive phase in order to reproduce successfully, and the average fraction of the reproductive span survived varies widely between groups: small rodents and birds in the wild survive on the average only 10 percent to 20 percent of their potential reproductive lifetimes; whales, elephants, apes, and other large mammals in the wild, on the other hand, live through 50 percent or more of their reproductive spans, and a few survive beyond reproductive age. In iteroparous forms the onset of senescence is gradual, with no evidence of specific systemic or environmental initiating mechanisms; senescence manifests itself early as a decline in reproductive performance. In species that grow to a fixed body size, decline of reproductive capacity begins quite early and accelerates with increasing age. In large egg-laying reptiles, which attain sexual maturity while relatively small in size and continue to grow during a long reproductive span, the number of eggs laid per year increases with age and body size but eventually levels off and declines. The reproductive span in such cases is shorter than the life-span.

These comparisons illustrate the influence exerted by factors of population dynamics on the evolution of reproductive and bodily (somatic) senescence. The proportional contribution of an individual to the rate of increase of the iteroparous population obviously diminishes as the number of his living progeny increases. In addition, his reproductive capacity diminishes with age. These facts imply that there is an optimum number of litters per lifetime. Whether or not these influences of population dynamics lead to the evolution of adaptive senescence patterns has long been debated by gerontologists but has not yet been investigated definitively.

Species differences in longevity and aging. That there are large differences in life-span between some species of animals has long been known, but only recently have the data become adequate for statistical analysis. Maximum life-span provides an estimate of the potential longevity of mammalian and avian species because of the sharp upper limit of the survival curves in life tables. Also, it is superior to the average life-span because the latter is influenced by environmental factors unrelated to aging (*e.g.*, human protection).

Maximum life-span as an indicator of longevity

The taxonomic stratification of longevity can be seen among the mammals. Primates, generally, are the longest lived group, although some small prosimians and New World monkeys have relatively short life-spans. The murid (mouselike) rodents are short-lived; the sciurid (squirrel-like) rodents, however, can reach ages two to three times longer than the murids. Three traits have independent correlations with life-span: brain weight, body weight, and resting metabolic rate. The dependence of life-span on these traits can be expressed in the form of an equation: $L = 5.5E^{0.54}S^{-0.34}M^{-0.42}$. Mammalian life-span ($L$) in months relates to brain weight (E) and body weight (S) in grams and to metabolic rate (M) in calories per gram per hour. The positive exponent for E (0.54) indicates that longevity of mammals has a strong positive association with brain size, independent of body size or metabolic rate. The negative coefficient for metabolic rate implies that life-span decreases as the rate of living increases, if brain and body weight are held constant. The negative partial coefficient for body weight indicates that the tendency for large animals to be longer lived results not from body size but rather from the high positive correlation of body weight with brain weight and its negative correlation with metabolic rate. The same kind of relation of L to E, S, and M holds for birds, but there is a tendency for birds to be longer lived than mammals of comparable brain and body size despite their higher body temperatures and metabolic rates. The larger reptiles have life-spans exceeding those of mammals of comparable size, but their rates of metabolism are about ten times lower, so that their total lifetime energy expenditures are lower than those for mammals. The more highly cephalized animals (*i.e.*, those with higher brain weight), especially the primates, have greater lifetime energy outputs; the total lifetime energy output per gram of tissue is about 1,200,000 calories for man and 400,000 calories for domestic animals such as cats and dogs.

The above relations hold for the homeothermic mammals, those with nearly constant body temperature. The heterothermic mammals, which are able to enter daily torpor, or seasonal hibernation, thereby reduce their metabolic rates more than tenfold. The insectivorous bats of temperate latitudes are the most dramatic example;

although they have life-spans in excess of 20 years, almost 80 percent of that time is spent in deep torpor. As a result, their lifetime energy expenditures are no greater than are those of other small mammals.

The longevities of arthropod species extend from a few days to several decades. The extremely short-lived insects have a brief single reproductive phase; the longer lived spiders and crustaceans are iteroparous, with annual reproductive cycles.

The inheritance of longevity. The inheritance of longevity in animal populations such as fruit flies and mice is determined by comparing the life tables of numerous inbred populations and some of their hybrids. The longevity of sample populations has been measured for more than 40 inbred strains of mice. Two experiments concur in finding that about 30 percent of longevity variation in female mice is genetically determined, whereas the heritability in male mice is about 20 percent. These values are comparable to the heritabilities of some physiological performances, such as lifetime egg or milk production, in domestic animals.

Rate of aging

The slope of the Gompertz function line indicates the rate of actuarial aging. The differences in longevity between species are the result primarily of differences in the rate of aging and are therefore expressed in differences in slope of the Gompertz function. Three species that vary in longevity, such as the opossum (seven years), rabbit (12 years), and cat (20 years), would have Gompertz functions like the three lines in B (see Figure 39), and they would have survival curves like the corresponding curves in 39A.

By courtesy of Argonne National Laboratory, Argonne, Illinois

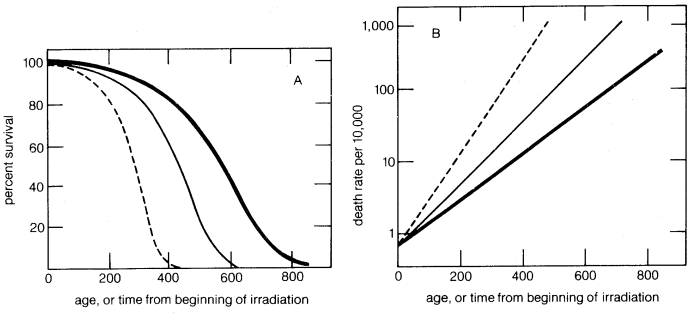


Figure 39: The relation of survivorship curves (A) to age-specific mortality rate curves (B). These curves also exemplify differences in the rate of aging and the accelerated aging due to continued exposure to ionizing radiation.

Comparison of life tables between mouse strains of a single species indicates that the strain differences result primarily from differences in age-independent hardiness factors. If strains differ in hardiness, the less hardy have death rates higher by a constant multiple at all ages, as shown by the parallel Gompertz functions. It is frequently found that the first-generation (F_1) hybrids of two inbred strains live longer than either parent. There has been no direct comparison of hybrid and inbred mice with regard to the rates of their biochemical aging processes, but life-table comparisons indicate that hybrid vigour is an increase of age-independent vigour and not a decrease in the rate of aging.

Recent research indicates that much of the variation in survival time between mouse strains is attributable to differences in inherited susceptibility to specific diseases. An important task of gerontology is to determine the extent of such genetic influences on aging.

The inheritance of longevity in humans is more difficult to investigate because length of life is influenced by socioeconomic and other environmental factors that generate spurious correlations between close relatives. A number of studies have been published, most of them pointing to some degree of heritability with regard to length of life or susceptibility to major diseases, such as cancer and heart disease. Although there is disagreement about the degree of heritability of longevity in man, the evidence for genetic transmission of susceptibility to coronary heart disease and related diseases is strong, as is the evidence that monozygotic (genetically identical) twins tend to have more similar life-spans than do like sex dizygotic (genetically different, fraternal) twins. (G.A.Sa.)

SENESCENCE IN PLANTS

The growth of the vascular plant depends upon the activity of meristems, which are, in a sense, always embryonic. Continued indefinitely, this mode of growth could mean immortality; indeed, the longest lived individual organisms ever to have existed on earth have been certain species of trees. Plants and plant parts, however, do die, and death is often not the consequence of accident or environmental stress but of physiological decline—aging, or senescence.

Various kinds of physiological senescence and death occur and may affect particular cells, tissues, organs, or the whole plant. In the formation of the vessels of the xylem, cells conclude their differentiation by dying and contribute their empty walls to the conducting tissue. Individual organs such as leaves usually have a limited life-span. Entire shoot systems may gradually die back in the aerial parts of perennial plants, which overwinter underground. And, finally, the whole plant may die after a limited period of growth and the completion of reproduction. This behaviour is found in many annual plants, which complete their life cycle in a single growing season. The life-span may extend to two years, as in biennial plants, or longer, as in banana and certain bamboos, which die after flowering and fruiting.

Programmed death of plant cells

In the examples cited above, the death of cells, organs, or individual plants appears to be "programmed" and, in some sense, adaptive. This is clearly so with the death of individual cells during differentiation, when residual products contribute to the effective function of the entire plant body. The death of leaves and of shoot systems is part of the plant's adaptation to the cycle of the seasons. In annual species, the death of the whole individual may be viewed in a similar way. The succession of generations in this case is carried on by seeds; the sacrifice of the parent plant may, in fact, contribute to the success of the seedling by making available to the seed a pool of reserves derived from the breakdown of parent tissues.

Certain features characterize the onset of senescence. The cells show degenerative changes often associated with the accumulation of breakdown products. Metabolic changes accompany the degeneration. Respiration may increase for a period, but the rate ultimately declines as the cellular apparatus degenerates. Synthesis of proteins and nucleic acids ceases, and, in some instances, disintegration of cells has been associated with the release of enzymes through the disruption of membrane-bounded bodies called lysosomes.

The death of individual cells in tissues such as the xylem appears to be governed by internal factors, but senescence often depends upon interaction of tissues and organs. The presence of young developing leaves often accelerates the aging of older leaves; removal of the younger leaves retards the senescence of the older ones, suggesting control by competition for nutrients. A similar effect is seen in annual plants, in which the development of fruits and seeds is associated with the senescence and, ultimately, the death of the rest of the plant; the removal of reproductive structures slows the rate of aging. In these instances competition obviously has some effect, but it does not sufficiently explain why older, mature organs suffer in competition with those still in active development. The link may lie partly in the capacity of developing organs to draw nutrients to themselves, even from older parts of the plant. Developing organs thus provide "sinks" toward which nutrients tend to move. The senescence of organs drained in this way could result from the progressive loss of certain key constituents; should leaf protein, for example, turn over by breakdown of proteins to their amino-acid constituents and then be resynthesized, a steady drain of amino acids from the leaf would progressively deplete the proteins in the leaf.

Nutritional "sinks"

"Sinks" can be only part of the explanation, however, for in detached leaves of plants such as tobacco, protein synthesis decreases, and protein content falls, while the amino-acid content actually rises. Senescence in such instances can hardly depend on the withdrawal of nutrients. Furthermore, leaf senescence can be retarded locally by the

application of cytokinins, hormones that stimulate plant cell division. Parallel effects have been demonstrated with growth substances of the auxin type in other plant systems. In the same way that active buds and fruits form sinks for nutrients from elsewhere in the plant, a cytokinin-treated area of a leaf attracts nutrients from other parts of the leaf. Although the metabolism of isolated leaves may differ in many respects from that of attached leaves, leaf senescence probably does not result only from nutrient drainage but also from the synthetic activity of leaf tissues, which may be under hormonal control from other parts of the plant. The root may be important, for roots are known to export cytokinins to the shoot.

Daily length of darkness

Environmental factors, primarily photoperiod (daily length of darkness) and temperature, play important parts in governing senescence and death in plants. In annual plants, death is the natural conclusion of development; thus, conditions accelerating development automatically advance senescence. This is readily seen in short-day plants, in which precocious reproduction upon exposure to long dark periods is followed by early death. Senescence may be retarded in these cases, however, by hormonal treatments of the kind known to delay degeneration and death in detached leaves. Competition for nutrients between vegetative and reproductive structures cannot be the primary cause of death, for, in species such as hemp, the male plants—which do not produce seeds—die earlier than the females under short-day (long-night) conditions.

In perennial plants, leaf fall is associated with approaching winter dormancy. In many trees leaf senescence is brought about by declining day length and falling temperature toward the end of the growing season. Chlorophyll, the green pigment in plants, is lost; yellow and orange pigments called carotenoids become more conspicuous; and, in some species, anthocyanin pigments accumulate. These changes are responsible for the autumn colours of leaves. There are some indications that day length may control leaf senescence in deciduous trees through its effect on hormone metabolism, for both gibberellins and auxins have been shown to retard leaf fall and to preserve the greenness of leaves under the short-day conditions of autumn.

From the foregoing it may be seen that senescence and death are important in the general economy of plants. The paradox that death contributes to survival is resolved when it is understood that the death of the part contributes to the better adaptation of the whole—whether organ, individual, or species. Viewed in this way, death is no more than another—albeit the ultimate—manifestation of development.

(J.H.-H.)

SENESCENCE IN MAMMALS

Changes in body composition, metabolism, and activity. The lean body mass, consisting of the skeletal muscles and all other cellular tissues, decreases steadily after physical maturity until, in extreme old age, it may be reduced to two-thirds its value in young adults. Body weight, however, usually increases with age, because stored fat and body water increases in excess of the loss of lean body mass. The relative amount of extracellular fluid increases with age during adult life, after decreasing steadily throughout fetal and postnatal development. Despite appearances, therefore, all tissues, even the skin, become more laden with water as a consequence of aging. The steady loss of voluntary (striated) muscle tissue mass throughout adult life depends somewhat on the pattern of physical activity. Evidence indicates that a large part of the loss of muscle mass with age is the result of disuse and atrophy rather than loss of muscle fibres.

The decrease of lean body mass is accompanied by a decrease in the level of overall metabolic activity. Basal metabolism is greatest during the period of most rapid mass growth; it then declines rapidly until physical maturity is reached and more slowly thereafter. In the rat the slow phase of decrease amounts to about 20 percent over a three-year period. The interior body temperature is maintained, despite lower heat production, by decreased blood flow through the skin with a consequent decrease of heat loss; the "cooling of the blood" with age, therefore, does not occur in the degree that might be inferred from the decrease in skin temperature. The amount of voluntary physical activity, such as running in an exercise wheel, typically decreases with age but varies considerably between individual animals.

Changes in structural tissues. The structural integrity of the vertebrate organism depends on two kinds of fibrous protein molecules, collagen and elastin. Collagen, which constitutes almost one-third of the body protein, is found in skin, bone, and tendons. When first synthesized by cells called fibroblasts, collagen is in a fragile and soluble form (tropocollagen). In time this soluble collagen changes to a more stable, insoluble form that can persist in tissues for most of an animal's life. The rate of collagen synthesis is high in youth and declines throughout life, so that the ratio of insoluble to soluble collagen increases with age. Insoluble collagen then builds up with age as a result of synthesis exceeding removal, much like another fibrous tissue, the crystalline lens of the eye. With increasing age, the number of cross-linkages within and between collagen molecules increases, leading to crystallinity and rigidity, which are reflected in a general body stiffness. There is also a decrease in the relative amount of a mucopolysaccharide (*i.e.*, the combination of a protein and a carbohydrate) ground substance; a measure of this, the hexosamine–collagen ratio, has been investigated as an index of individual differences in the rate of aging. An important consequence of these changes is decreased permeability of the tissues to dissolved nutrients, hormones, and antibody molecules.

The rate of aging of collagen is related to the overall metabolic activity of the animal; rats kept on low calorie diets have more youthful collagen than fully nourished rats of the same age.

Elastin is the molecule responsible for the elasticity of blood vessel walls. With age, progressive loss of elasticity of vessels occurs, presumably because of fragmentation of the elastin molecule.

Chemical retardation of aging

The cross-linkage of collagen is chemically similar to the cross-linkages that occur in skins when they are tanned to leather. This similarity has stimulated proposals that chemicals that inhibit cross-linkage in tanning will retard aging. Such compounds have been tried on animals, but the problems of toxicity have not yet been solved.

Tissue cell loss and replacement. The tissues of the body fall into two groups, according to whether or not there is continuous renewal of tissue cells. At one extreme are nonrenewal tissues such as nerves and voluntary muscles, in which no new cells are formed (at least in mammals) after a certain stage of growth. In renewal tissues such as the intestinal epithelium and the blood, on the other hand, some cell types live only one or a few days and must be replaced hundreds of times in the life-span of even a short-lived animal such as the rat. Between these limits lie many organs, such as liver, skin, and endocrine organs, that have cells that are replaced over periods ranging from a few weeks to several years in man.

A peripheral nerve is a convenient object to study because the total number of fibres in the nerve trunk can be counted. This has been done for the cervical and thoracic spinal nerve roots of the rat, the cat, and man. In the ventral and dorsal spinal roots of man, the number of nerve fibres decreases about 20 percent from age 30 to age 90. In the cat, the rat, and the mouse, however, the data do not consistently indicate a decrease of number of spinal root fibres with age. In man the number of olfactory nerve fibres, which serve the sense of smell, decreases by age 90 to about 25 percent of the number present at birth, and the number of optic nerve fibres, serving vision, decreases at a nearly comparable rate.

There is a striking decrease in the number of living cells in the cerebral cortex of the brain of humans with age. The cerebellar cortex of the rat and man is about as susceptible to age deterioration as is the cerebral cortex. Other parts of the brain are not so obviously marked by aging.

There is, in short, a tendency for the higher and more recently evolved levels of the nervous system to undergo more severe aging loss than do other regions, such as the brain stem and spinal cord. It is not yet known how much

of the loss of brain cells results from conditions within the brain itself and how much results from extrinsic causes, such as deterioration of the blood circulation. The nutrition and maintenance of nerve cells, or neurons, in the central nervous system depends to a considerable extent on neuroglia, small cells that surround the neurons. The absolute number of these cells apparently does not decrease with age, but some of the microscopic changes seen in the neurons of old persons are similar to the changes produced by starvation or physical exhaustion.

It has been shown that after an attack of measles, the virus remains in the host's body for the remainder of life and infrequently gives rise to a rapidly progressing degeneration of the cerebral cortex. This virus or other inapparent viruses may also be responsible for the individual differences in onset of senility in man.

The renewal tissues are typically made up of a population of proliferative cells, which retain the capability for division, and a population of mature cells, produced by the proliferative cells and with limited life-spans. The production of cells must balance the steady loss and also compensate quickly for unusual losses caused by injury or disease, so each renewal tissue has one or more channels of feedback control to adjust production to demand. Aging of renewal tissues is expressed in several ways, including decrease in the number of proliferative cells, decrease in the rate of cell division, and decrease in responsiveness to feedback signals. Changes of these factors in the blood-forming tissues of the mouse are small, yet the blood-forming tissues do suffer an aging deficit, for the ability to respond to extreme or repeated demand is significantly reduced in older mice.

Senescence of renewal tissues

The intact skin has a cell turnover time of several weeks, with the capability, shared by all renewal tissues, of temporarily increasing the rate of cell production by a large factor in response to injury. The rate of wound healing decreases with age, rapidly at first and more slowly as age increases.

One of the most regular and striking aging processes is the decrease in the ability to focus on both close and distant objects. This loss in visual accommodation is the result in part of a weakening of the ciliary muscle of the eye and of a decrease in the flexibility of the lens. A further contributing factor, however, is that the lens continues to grow throughout life at a rate that diminishes with age. This growth is the result of continuous division of epithelial cells near an imaginary midline of the lens, giving rise to fresh cells that differentiate into the precisely aligned lens fibres. Once formed, the fibres remain permanently in place.

An important feature of the renewal mechanism is the stem cell. These cells, which may normally continue to divide at a low rate throughout life, under conditions of increased demand enter a compensatory proliferative phase during which they divide rapidly. Blood-forming tissue has a stem-cell population that responds to injury readily in youth, but its capacity diminishes with age. The increased incidence of anemia in old age and the reduced capacity to respond to blood loss have been attributed to depletion of the blood-forming stem cells. Stem-cell populations have not been identified with certainty in other proliferative tissues. The intestinal mucosa, in particular, has a high cell-division rate without any clear indication of a reserve population of stem cells.

Mammalian cell cultures. Dividing cells from various mammalian tissues can be grown in vitro (outside the body) under careful laboratory control. Various lines of cancer cells have been grown in continuous culture for many decades. In the early period of tissue-culture technology it was claimed that certain chicken cells (fibroblasts) had been maintained in culture for 20 years. This led to the belief that dividing cells were potentially immortal and focussed interest on nondividing cells as the seat of the aging process. This view has lost standing in recent years. It has now been established that a population (clone) of fibroblasts has a finite life history in culture. It has a period of healthy growth, during which it can be transferred, or "split," several dozen times, indicating that the cells have undergone more than that number of generations. The cultures, however, go into a senescent phase and die out, usually before the 50th transfer. Occasionally, the chromosomes in a cell in the culture undergo a mutation (change) that results in a loss of a growth-limiting factor, leading to the establishment of a subclone capable of indefinite growth. This happens fairly often in cultures of mouse cell strains but only rarely in cultures of human cells. Such mutations usually involve chromosomal rearrangements or changes in the number of chromosomes.

The present view, therefore, is that dividing mammalian cells with a normal chromosomal complement have a limited growth potential and that the capacity for indefinite growth shown by cancer cells and transformed cells is the result of the loss of a growth-limiting factor. The number of transfers that cell strains can undergo decreases as the age of the donor increases, in a way reminiscent of the decreased turnover rate of fibroblasts in living chickens and of the decreased rate of wound healing with age.

Changes in tissue and cell morphology. There are numerous instances of tissue changes with age. The atrophy of tissues of moderate degree is usual. The shrinkage of the thymus is especially striking and important in view of its role in immunological defense. The diminution of cellular tissue and replacement by fatty or connective tissue is prominent in marrow and skin. In the kidney, entire secretory structures (nephrons) are lost. The secretory cells of the pancreas, thyroid, and similar organs decrease in numbers.

Wasting of tissues with age

An important age change is the accumulation of pigments and inert—possibly deleterious—materials within and between cells. The pigment lipofuscin accumulates within heart muscle cells; it is not detectable at ten years of age but rises to almost 3 percent of the cell volume by age 90. Amyloid substance, a protein–carbohydrate complex, increases in tissues in middle age; it is presumably a product of autoimmune reactions, immune reactions misdirected against the organism itself. In an extreme case of a rare autoimmune disease, amyloid disease, particular organs are virtually choked with amyloid substance. Trace metals also accumulate in various tissues with age, and although the amounts are very small, certain metals can poison enzyme systems, stimulate mutations, or cause cancer.

AGING AT THE MOLECULAR AND CELLULAR LEVELS

Aging of genetic information systems. The physical basis of aging is either the cumulative loss and disorganization of important large molecules (*e.g.,* proteins and nucleic acids) of the body or the accumulation of abnormal products in cells or tissues. A major effort in aging research has been focussed on two objectives: to characterize the molecular disruptions of aging and to determine if one particular kind is primarily responsible for the observed rate and course of senescence; and to identify the chemical or physical reactions responsible for the age-related degradation of large molecules that have either informational or structural roles. The working molecules of the body, such as enzymes and contractile proteins, which have short turnover times, are not thought to be sites of primary aging damage. The deoxyribonucleic acid (DNA) molecules of the chromosomes appear to be potential sites of primary damage, because damage to DNA corrupts the genetic message on which the development and function of the organism depend. Damage at a single point in the DNA molecule can be followed by the synthesis of an incorrect protein molecule, which may result in the malfunction or death of the host cell or even of the entire organism. Attention therefore has been given to the somatic mutation hypothesis, which asserts that aging is the result of an accumulation of mutations in the DNA of somatic (body) cells. Aneuploidy, the occurrence of cells with more or less than the correct (euploid) complement of chromosomes, is especially common. The frequency of aneuploid cells in human females increases from 3 percent at age 10 to 13 percent at age 70. Each DNA molecule consists of two complementary strands coiled around each other in a double helix configuration. Evidence indicates that breaks of the individual strands occur with a higher frequency than was once suspected and that virtually all such breaks are repaired by an enzymatic mechanism that

DNA damage

destroys the damaged region and then resynthesizes the excised portion, using the corresponding segment of the complementary strand as a model. The mutation rate for a species is therefore governed more by the competence of its repair mechanism than by the rate at which breaks occur. This may help to explain why the mutation rates of different species are roughly proportional to their generation times and justifies research to determine whether the enzymatic mechanisms involved are accessible to control. It remains to be seen whether a reduction of mutation rates will retard the onset of generalized aging or of a specific disease process.

There are, however, serious objections to the somatic mutation theory. The wasp *Habrobracon* is an insect that reproduces parthenogenetically (*i.e.,* without the need of sperm to fertilize the egg). It is possible to obtain individuals with either a diploid, or paired, set of chromosomes, as in most higher organisms, or a haploid, single, set. Any gene mutation in a haploid cell at an essential position would result in loss of a vital process and impairment or death of the cell; in a diploid cell a serious mutation is often compensated for by the complementary gene and the cell can carry on its vital functions. Experiments have shown that haploid wasps live about as long as diploids, implying either that mutations are not a quantitatively important factor in aging or that parthenogenetic species have compensated for the vulnerability of their haploids by developing an increased effectiveness of DNA repair.

Chromosomes can be separated into DNA and protein molecules, but with increasing difficulty in older cells. The isolated DNA of old animals, however, does not differ from that of the young. Although most of the DNA in a given cell at a given time is repressed (*i.e.,* blocked from functioning), it is more repressed in old animals; it is not yet known whether this is a primary age change or a consequence of reduced cell metabolism arising from other causes.

Aging of the immunological system. Another important molecular information system of the body is the immunological system, part of which, the thymus-dependent subsystem, is specialized for defense against invading microorganisms and for the as yet poorly understood role of detecting and removing body cells that have changed in such ways that they are no longer recognized by the body as part of its own substance, leading to the autoimmune reactions mentioned above. The immunological system has been implicated in the body's defenses against cancer. Cancerous growths (neoplasms) are thought to arise from single cells that undergo a drastic transformation as a result of either a genetic mutation or the activation of a latent (hidden) virus that may have been transmitted genetically from parent to offspring. The control of cancer susceptibility by genetically governed defense mechanisms has been indicated by the breeding of high and low cancer susceptibility in mice. There is a growing body of evidence that the thymus-dependent immunological system is instrumental in repressing the development of cancer.

Aging and cancer

One piece of evidence is that the immunosuppressive procedures of organ transplantation are often followed by a greatly increased incidence of neoplasms. The thymus-dependent system can itself, however, give rise to age-related autoimmune disease, in which the immunological system perceives normal body tissue as foreign and attacks it with antibodies. The initial step in these diseases is considered to be a somatic mutation in a single cell of the immunological system. Such considerations are the basis of several immunological theories of aging, which seek to explain the phenomena of senescence in terms of mutations in the immunological system.

Aging of neural and endocrine systems. The loss of psychological and neurophysiological capacities with age is undoubtedly the result, in large part, of the loss of neurons, but deficiencies in the metabolic processes of the surviving cells are demonstrably involved. The ability of the eye to dark-adapt (*i.e.,* increase its sensitivity at low light levels) decreases with age, but part of that decrease can be restored by breathing pure oxygen. Various mental processes in old people are also found to be improved by breathing oxygen. The establishment of a memory trace (connections in the brain that are associated with memory) involves the synthesis of protein; any slowed induction of protein synthesis, as from lower oxygen intake, with age could be a factor in the deficits of learning and memory of old people.

A general characteristic of aging of the endocrine system is that the cells that once responded vigorously to hormones become less responsive. A normal chemical in cells, cyclic adenosine monophosphate (AMP), is thought to be a transmitter of hormonal information across cell membrane; it may be possible to identify the specific sites in the membrane or the cell interior at which communication breaks down.

INTERNAL AND EXTERNAL CAUSES OF AGING

External environmental agents. *Ionizing radiations.* The shortening of life caused by ionizing radiations (*e.g.,* X-rays) has been determined for many species, including mice, rats, hamsters, guinea pigs, and dogs. The occurrence of some diseases, such as leukemia, may increase disproportionately after irradiation, with the degree of increase influenced by age and sex.

The permanent nature of radiation damage is shown by the comparison of life-spans of irradiated and control populations. An irradiated population dies out like a chronologically older unirradiated population. Members of a population given a single dose of X-rays or gamma rays in early adult life die of the same diseases that afflict the unirradiated control population, but they die months or even years earlier.

Continuous irradiation throughout life at low dose rates (daily doses from one-thousandth to one-tenth the dose that would kill immediately) speeds the mortality process. It is not yet clear if the molecular damage produced by such irradiation is the same as the molecular changes that accompany natural aging. Studies of animals and of cells grown in culture suggest that large doses of radiation kill by producing deleterious rearrangements of chromosomes in the proliferative cell population. Such aberrations also increase with age, but they seem to be less important in the natural aging process. At low radiation doses, chromosome aberrations become relatively less important than other effects, and the primary radiation damage in these conditions may bear a closer relation to the aging lesion. Under conditions of low-dose irradiation, however, the only definite effect is slightly increased cancer incidence; a generalized aging effect has not yet been observed.

The aging effect of natural radiation

Natural radioactivity in the body, arising mostly from radioactive potassium and radium, and natural background irradiation, from the Earth and from cosmic rays, are not major contributors to the aging process, even in the long-lived human species. They are responsible, however, for a small percentage of cancer incidence. Although the dose to the body from medical radiations is a fraction of the background level and the radiation from nuclear weapon test fallout is less than 1 percent of the background, both sources contribute to cancer induction in proportion to their amounts.

Temperature. Flour beetles, fruit flies, fishes, and other poikilothermic (temperature-variable) organisms live longer at the lower range of environmental temperature. These observations led to the rate-of-living hypothesis, which, simply stated, holds that an organism's life-span is dependent on some critical substance that is exhausted more rapidly at higher temperature. Careful analysis of the data on temperature–longevity relations shows, however, that the rate-of-living hypothesis is inadequate in its original form. The most telling evidence comes from experiments in which fruit flies were kept at one temperature for part of their lives and at another temperature for the remainder. The results are not consistent with the rate-of-living hypothesis, but no satisfactory theory has appeared as yet to take its place. An important factor that has not yet been adequately taken into account is the relation of metabolic efficiency to temperature. The energy cost of the biosynthetic processes studied has been discovered to be minimal at an intermediate temperature in the range to which the species is adapted and to increase at higher or lower temperatures. A related phenomenon holds for

longevity; the number of calories expended by fruit flies per lifetime is maximal at an intermediate temperature, so the rate of aging per calorie is minimal at that temperature.

There is a question of the degree to which aging occurs as a result of heat destruction (thermal denaturation) of proteins. Thermal denaturation is predominately a disruption of the folding of molecules, which requires the breaking of numbers of low-energy bonds. It seems not to be a strong contributing factor to aging. There is still the possibility that rare events, such as mutations, may arise to a significant degree from thermal denaturation.

Physical wear of nonrenewable structures. One of an animal's most important assets is its chewing apparatus, including jaws and teeth. Adaptation to tooth rate of wear is especially important for animals that consume large quantities of grass and herbage. Such adaptations include higher tooth crowns (hypsodonty), larger grinding area, and longer tooth growth period. Tooth wear may be limiting for survival in adverse environments, but, on the whole, it is not an important life-limiting characteristic. The same can be said for other external organs subject to physical wear.

Infectious disease and nutrition. The populations in poor environments, characterized by high rates of infectious disease and poor nutrition, have higher death rates than populations in good environments at all ages, yet there is no positive evidence that disadvantaged populations experience a higher rate of aging.

Rats kept on diets restricted in calories live longer and have lower cancer incidence than do rats that are allowed to eat at will; maximum longevity, however, is achieved at a nutritional level that keeps the animal sexually immature and below normal weight.

Internal environment: consequences of metabolism. The metabolic activities of organisms produce highly reactive chemicals, including strong oxidizing agents. The internal structure of the cell, however, minimizes the harmful effects of such agents; the critical reactions take place within enclosed structures such as ribosomes, membranes, or mitochondria, and counteractive enzymes such as peroxidases are present in abundance. It is nevertheless likely that low concentrations of these reactive substances can reach vital molecules and contribute to the characteristic rate of aging injury. Experiments in which mice are fed low levels of antioxidants such as butylated hydroxytoluene (BHT) have been encouraging but are still somewhat equivocal.

Aging and membrane integrity

Membranes are the site of much of the metabolic activity of cells; they provide the barriers that keep incompatible reactions separated. Certain membrane structures, called lysosomes, contain enzymes capable of digesting the cell if released; the stability of cells and organisms is therefore very much bound up with the stability of membranes. A number of drugs, including corticosteroids, salicylates, and antihistamines, act by stabilizing cell membranes against inflammatory stimuli. Some of them are found to prolong life in fruit flies and to prolong survival of cells in vitro. The mode of action of these drugs is connected to substances called prostaglandins, which can alter specific membrane characteristics. (G.A.Sa.)

Human aging

Aging begins as soon as adulthood is reached and is as much a part of human life as are infancy, childhood, and adolescence. Gerontology (the study of aging) is concerned primarily with the changes that occur between the attainment of maturity and the death of the individual. The goal of research in gerontology is to identify the factors that influence these changes. Application of this knowledge is expected to reduce the disabilities now associated with aging.

Human aging has many aspects, both for the individual and for society, but the primary ones fall into three major categories; namely, biological-physiological, psychological-behavioral, and social-economic.

The biological-physiological aspects of aging include both the basic biological factors that underlie aging and the general health status. Since the probability of death increases rapidly with advancing age, it is clear that changes must occur in the individual which make him more and more vulnerable to disease. For example, a young adult may rapidly recover from pneumonia, whereas an elderly person may die. (See above *Biological theories of aging.*)

Physiologists have found that the performance of many organs such as the heart, kidneys, brain, or lungs shows a gradual decline over the life-span. Part of this decline is due to a loss of cells from these organs, with resultant reduction in the reserve capacities of the individual. Furthermore, the cells remaining in the elderly individual may not perform as well as those in the young. Certain cellular enzymes may be less active, and thus more time may be required to carry out chemical reactions. Ultimately the cell may die.

EFFECT OF AGING OF THE BODY SYSTEMS

Heart disease

Cardiovascular system. Diseases of the heart are the single largest cause of death after age 65. Thus, with increasing age the heart becomes more vulnerable to disease. Even in the absence of detectable disease, the heart undergoes deleterious changes with advancing age. Structural changes include a gradual loss of muscle fibres with an infiltration of fat and connective tissue. There is a gradual accumulation of insoluble granular material (lipofuscin, or "age pigment") in cardiac muscle fibres. These granules, composed of protein and lipid (fat), make their first appearance by the age of 20 and increase gradually, so that by the age of 80 they may occupy as much as 5–10 percent of the volume of a muscle fibre.

The heart also shows a gradual reduction in performance with advancing age. The amount of blood pumped by the heart diminishes by about 50 percent between the ages of 20 and 90 years. There are marked individual differences in the effects of age. For example, some 80-year-old individuals may have cardiac function that is as good as that of the average 40-year-old individual.

Under resting conditions, the heart rate does not change significantly with age. During each beat, however, the muscle fibres of the heart do not contract as rapidly in the old as in the young. This reduction in power, or rate of work, is due to the age-associated reduction in the activities of certain cellular enzymes that produce the energy required for muscular contraction.

In spite of these changes, the heart, in the absence of disease, is able to meet the demands placed upon it. In response to physical exercise it can increase its rate to double or triple the amount of blood pumped each minute, although the maximum possible output falls, and the reserve capacity of the heart diminishes with age.

Blood vessels and blood pressure

Arteriosclerosis, or hardening of the arteries, increases markedly in incidence with age, and is often regarded as part of aging. This is not necessarily true. Arteriosclerosis may appear even in adolescents. It is a progressive disorder and is present to some extent in practically all individuals by middle life. It is, therefore, impossible to make a clear distinction between the effects of aging and the effects of disease in blood vessels in human beings. In some animal species, as, for example, the rat, that do not develop arteriosclerosis, age changes in the heart and blood vessels can be identified.

In general, blood vessels become less elastic with advancing age. There is a progressive thickening of the walls of larger blood vessels with an increase in connective tissue. The connective tissue itself becomes stiffer with increasing age. This occurs because of the formation of cross-links both within the molecules of collagen, a primary constituent of connective tissue, and between adjacent collagen fibres. These changes in blood vessels occur even in the absence of the deposits on the arterial wall characteristic of atherosclerosis, which interfere with blood flow through the arteries. The gradual loss of elasticity increases with resistance to the flow of blood so that blood pressure may increase. This in turn increases the work that the heart must do in order to maintain the flow of blood.

While both systolic and diastolic blood pressures (blood pressures at contraction and dilation of the heart, respectively) increase with age, the rate of systolic increase exceeds that of diastolic so that the pulse pressure widens (see Table 5). The increase in pressure stops in the eighth

Table 5: Mean Blood Pressure in Adults by Age and Sex, 1960–62
(millimetres of mercury)

age (years)	men		women	
	systolic	diastolic	systolic	diastolic
18–24	121.7	71.6	111.8	69.4
25–34	124.7	76.4	115.6	72.9
35–44	128.6	80.7	122.8	78.0
45–54	133.8	83.2	133.8	82.0
55–64	140.3	83.1	146.6	84.9
65–74	148.0	81.0	160.2	83.7
75–79	154.3	79.4	156.6	79.3

Source: U.S. Public Health Service.

decade of life, and there may even be a slight decline in pressure in extreme old age.

On the average, obese people have higher blood pressures than those with normal body weights. Since the incidence of obesity increases with age at least up to the age of 55–60, this factor may contribute in part to the increase in blood pressure with age.

Digestive system. Loss of teeth, which is often seen in elderly people, is more apt to be the result of long-term neglect than a result of aging itself. The loss of teeth and incidence of oral disease increase with age, but, as programs of water fluoridation are expanded and the incidence of tooth decay in children is reduced, subsequent generations of the elderly will undoubtedly have better teeth than the present generation.

While it is true that the secretion by the stomach of hydrochloric acid, as well as other digestive enzymes, decreases with age, the overall process of digestion is not significantly impaired in the elderly. Sugar, proteins, vitamins, and minerals are absorbed from the stomach and intestine as well in the elderly as in the young. Some investigations indicate a slight impairment in fat absorption, but the reduction is probably of little practical significance.

These findings have important implications for nutrition of the elderly. There is no evidence that the intake of any nutrient, such as vitamins and minerals, need be increased in the elderly because of impaired absorption. Nutritional deficiencies can be avoided as long as the diet is varied to assure adequate intake of all nutritional elements. Deficiencies are most likely to develop from poor eating habits, such as excessive intake of carbohydrate with a reduction in protein. In the elderly these deficiencies are most apt to be in the intake of protein, calcium, iron, vitamin A, and thiamine (also called vitamin B_1).

Nervous system. Changes in the structures of the brain due to normal aging are not striking. It is true that with advancing age there is a slight loss of neurons (nerve cells) in the brain. This is because, in the adult, neurons have lost the capacity to form new neurons by division. The basic number of neurons in the brain appears to be fixed by about the age of 10. The total number of neurons is extremely large, however, so that any losses probably have only a minor effect on behaviour. Since the physiological basis of memory is still unknown, it cannot be assumed that the loss of memory observed in elderly people is caused by the loss of neurons in the brain.

Neurons are extremely sensitive to oxygen deficiency. Consequently, it is probable that neuron loss, as well as other abnormalities observed in aging brains, results not from aging itself, but from disease, such as arteriosclerosis, that reduces the oxygen available to areas of the brain by reducing the blood supply.

Changes in the brain

There are probably functional changes in the brain that account for the slowing of responses and for the memory defects that are often seen in the elderly; and even small changes in the connections between cells of the brain could serve as the basis for marked behavioral changes, but, until more is known about how the brain works, behavioral changes cannot be related to physiological or structural changes. It is known that, because of the slow course of aging, the nervous system can compensate and maintain adequate function even in centenarians.

Human behaviour is highly dependent on the reception and integration of information derived from sensory organs, such as the eye and ear, as well as from nerve endings in skin, muscle, joints, and internal organs. There is, however, no direct relation between the sensitivity of receptors and the adequacy of behaviour, because the usual level of stimulation is considerably greater than the minimum required for stimulation of the sense organs. In addition, an individual adapts to gradual impairments in one sensory organ by using information available from other sense organs. Modern technology has also provided glasses and hearing aids to compensate for reduced acuity in the sense organs.

The incidence of gross sensory impairments, of which many are the result of disease processes, increases with age. One survey conducted in the United States classified 25.9 per 1,000 persons aged 65–74 as blind, in contrast to 1.3 per 1,000 aged 20–44 years. In the age group 65–74, 54.7 per 1,000 persons were classified as functionally deaf, compared with 5.0 per 1,000 in the age range 25–34 years.

Vision. Visual acuity (ability to discriminate fine detail) is relatively poor in young children and improves up to young adulthood. From about the middle 20s to the 50s there is a slight decline in visual acuity, and there is a somewhat accelerated decline thereafter. This decline is readily compensated for by the use of eyeglasses. There is also reduction in the size of the pupil with age. Consequently, vision in older people can be significantly improved by an increase in the level of illumination.

Aging also brings about a reduction in the ability to change the focus of the eye for viewing near and far objects (presbyopia), so that distant objects can ordinarily be seen more clearly than those close at hand. This change in vision is related to a gradual increase in rigidity of the lens of the eye that takes place primarily between the ages of 10 and 55 years. After age 55 there is little further change. Many people in their 50s adopt bifocal glasses to compensate for this physiological change.

The sensitivity of the eye under conditions of low illumination is less in the old than in the young; that is, "night vision" is reduced. Sensitivity to glare is also greater in the old than in the young.

The incidence of diseases of the eye, such as glaucoma and cataracts (characterized, respectively, by increased intra-ocular pressure and opaque lenses), increases with age, but recent advances in surgery and the development of contact lenses have made it possible to remove cataracts and restore vision to many individuals.

Hearing. Hearing does not change much with age for tones of frequencies usually encountered in daily life. Above the age of 50, however, there is a gradual reduction in the ability to perceive tones at higher frequencies. Few persons over the age of 65 can hear tones with a frequency of 10,000 cycles per second. This loss of perception of high frequencies interferes with identifying individuals by their voices and with understanding conversation in a group, but does not ordinarily represent a serious limitation to the individual in daily life. Listening habits and intellectual level play an important role in determining the ability to understand speech, so that there is often a disparity between measurements of pure tone thresholds and ability to perceive speech.

Loss of perception of high frequencies

Other sensory impairment. After the age of 70 other sense organs may show a reduction in sensitivity. Reduced taste sensitivity is associated with atrophy and loss of taste buds from the tongue in the elderly. The effect of aging on the sense of smell has not been precisely determined because this sense is extremely difficult to assess quantitatively; in addition, smoking and exposure to occupational odours and noxious substances in the air influence sensitivity to smells.

Sensitivity to pain is difficult to evaluate quantitatively under controlled laboratory conditions. There is some evidence that it diminishes slightly after the age of 70.

There is a general slowing of responses in the elderly. Reflexes become slightly more sluggish and the speed of conduction of impulses in nerves is slightly slowed. Old people require more time to respond to the appearance of a light than do young. The slowing with age is greater in situations where a decision must be made. For example,

more time is required to initiate a response in experiments in which the instructions are "Press the button with your right hand when the green light comes on, but with your left hand when the red light comes on" than if the instructions are, "Push the button if either light comes on." From these and other experiments it is concluded that the primary site of slowing of responses is within the brain rather than in the end organ (eye) itself.

Skin. The primary age change in the skin is a gradual loss of elasticity. Although this basic change plays a role, other factors, such as exposure to the weather and familial traits, also contribute to the development of wrinkles and the pigmentation associated with senescence. The ability of the skin to take up slack and remain closely adherent to the underlying structures is due to the presence of fibres of the proteins elastin and collagen. Studies of the minute structures of the skin show a gradual reduction in elastin. In addition, the collagen fibres show an increase in cross-links, which greatly restricts the elastic properties of the collagen network.

The effectiveness of facial massage in retarding the development of wrinkles has not been evaluated under carefully controlled conditions. The application of creams containing female sex hormone stimulates regeneration of skin and improves its elastic properties. Other effects, which may be undesirable or even hazardous, may follow repeated administration of these hormones.

Thyroid and adrenal hormones

Endocrine system. Because of the importance of hormones in the regulation of many physiological systems, impairments in endocrine (ductless) glands have traditionally been cited as important determinants in aging.

Thyroxine, the hormone secreted by the thyroid gland, regulates the level of activity of all the cells of the body. When thyroxine secretion is reduced, all metabolic processes proceed at a reduced rate and basal metabolism falls. (Metabolism consists of the chemical changes taking place within the cells of an organism during the processes of growth and restoration of tissues and the production of energy necessary for bodily processes; basal metabolism is the metabolism, as measured by the rate at which heat is given off, when an organism is in a resting and fasting state.) Since basal metabolism decreases with age, it seemed reasonable to ascribe aging to a loss of thyroid function, but this assumption has proved to be incorrect. Experimental studies have shown that the ability of the thyroid gland to produce thyroxine is not reduced in the elderly, and that there is a reduction in the utilization of thyroxine in various tissues of the body. Further studies of cellular metabolism are needed to find out why this is so.

Since aging is associated with reduced ability to adjust to stresses, and since the adrenal cortex (the outer part of the adrenal gland) plays a role in many of these adjustments, numerous attempts have been made to assess senescent changes in the function of the adrenal cortex. Although after the age of 50 there is a reduction in blood levels of the hormones secreted by the adrenal cortex, the ability of the gland to produce hormones when stimulated by the experimental administration of adrenocorticotrophic hormone (ACTH), the pituitary hormone that regulates the activity of the adrenal cortex, has been shown to be as good in the old as in the young.

Hormones of pituitary and of pancreas

The pituitary gland is often referred to as the master gland of the body, since it produces hormones that stimulate the activities of other endocrine glands, such as the adrenal, the thyroid, and the ovary. It was therefore once assumed that reduction in the function of these glands associated with aging is due to lack of proper stimulation from the pituitary gland. Methods for determination of the very small amounts of these regulating hormones present in the blood have been developed and as yet no systematic studies of age differences in blood levels of these hormones have been reported.

The pancreas secretes insulin, the hormone that regulates the utilization of sugar and other nutrients in the body. When the pancreas fails to produce adequate amounts of insulin, diabetes occurs. One test for diabetes involves measuring the rate of removal of sugar from the blood, that is, the glucose-tolerance test. One characteristic of aging is a reduction in the rate of removal of excess sugar from the blood. At present it is not known whether this represents the early stages of diabetes or whether it is a normal age change. It does appear in aged individuals who do not show any of the other symptoms of diabetes. Furthermore, it has been shown that, unlike the diabetic, elderly subjects can, with additional stimulation, produce more insulin. In normal young persons the pancreas releases more insulin in response to even a slight rise in blood sugar levels. In the elderly, the sensitivity of the pancreas is reduced so that a higher level of blood sugar is required to stimulate it to action. With maximum stimulation the pancreas in the aged can produce as much insulin as the pancreas in the young.

Sexual activity

It has long been known that the excretion of both male and female sex hormones diminishes with age. In the female, the excretion of estrogens (female sex hormones) falls markedly at the menopause (see REPRODUCTION AND REPRODUCTIVE SYSTEMS: *The human reproductive system: The female reproductive system: Menopause*). In the male, the excretion of androgens (male sex hormones and their degradation products) falls gradually over the age span 50–90, so that the existence of a male "climacteric" is highly improbable.

Sexual activity, as reported in interview studies, diminishes progressively between the ages of 20 and 60 in both males and females. In males the frequency of marital intercourse falls from an average of four per week in 20-year-olds to one per week in 60-year-olds. Practically all males aged 20–45 reported some level of sexual activity. Between the ages of 45 and 60 only about 5 percent of males reported loss of sexual activity.

Few systematic studies have been made of sexual behaviour in individuals over the age of 60, but clinical reports indicate that at least some males remain sexually active at 90.

There are wide individual differences in the level of sexual activity in both males and females. In human beings, sexual behaviour is influenced more by psychological and social factors than by the levels of sex hormones circulating in the blood. Nevertheless, the use of male sex hormones has had a long, and stormy, history as a rejuvenating agent for males. Attempts to rejuvenate elderly males by injecting crude extracts from testicles of animals, as well as various androgenic compounds, were made, but the effects, if any, were only transitory. In the early 1900s, sex glands from other animals were transplanted into human beings, but the results were questionable and the side effects were often disastrous. At about the same time, an operation was devised in which the spermatic ducts were tied off. It was assumed that preventing the loss of sperm would stimulate the sex glands to produce androgenic hormones which would rejuvenate the individual. None of these assumptions proved correct, so that the operation was soon abandoned as a rejuvenating procedure.

Since tissue loss does occur with aging, the administration of anabolic hormones (hormones that promote the buildup of tissues) may represent an important future development. The compounds that are currently available have a number of undesirable side effects and cannot be used routinely. Chemists and pharmacologists continue research to produce new steroids that will have anabolic effects without the undesirable side effects.

Calcium loss

Skeletal system. With aging, the bones gradually lose calcium. As a result they become more fragile and are more likely to break, even with minor falls. Healing of fractures is also slower in the old than in the young. Recent advances in orthopedic surgery, with the replacement of parts of a broken bone or joint with new structures or the introduction of metallic pegs to hold broken parts together, have been of great value to elderly people.

The incidence of osteoporosis, a disease characterized by a loss of calcium and minerals from bone, also increases with age. It occurs more frequently in women after menopause than in men and is especially evident in the spinal column. Back pain is a primary symptom of the disease. It can be treated by increasing calcium intake in association with the administration of anabolic hormones.

The mobility of joints diminishes with age and the incidence of arthritis increases.

Respiratory system. Vital capacity, or the total amount of air that can be expelled from the lung after a maximum inspiration, diminishes with age, as does the total volume of air that can be contained in the lungs. In contrast, the amount of air that cannot be expelled from the lung increases. These changes in respiratory mechanisms are primarily a reflection of the increased stiffness of the bony cage of the chest and decreased strength of the muscles that move the chest during respiration.

The lung also contains elastin and collagen to give it elastic properties. As indicated previously, the formation of cross-links in elastin and collagen that takes place with aging reduces the elastic properties of the lung.

The transfer of oxygen and carbon dioxide from the air in the lungs to the blood is influenced by the amount of blood flowing through the lungs as well as by the amount of air moved in and out. The characteristics of the membranes that separate blood and air in the lungs are also important in maintaining an adequate supply of oxygen to the body. Although with age there is a slight reduction in the amount of oxygen that can be moved from the air to the blood in the lungs, the reduction becomes apparent only when large amounts of oxygen are required, as during strenuous exercise. It is believed that a primary factor in the impairment of oxygen transfer in the lungs of elderly subjects is the lack of appropriate adjustment of the blood flow to the air sacs in the lung.

ncidence of emphysema

Emphysema, abnormal distension of the lungs with air, is a lung disease reaching its highest incidence between the ages of 45 and 65. In the United States the death rate from emphysema increased by almost 400 percent between 1950 and 1960. Although the exact causes of the disease are still unknown, the presence of noxious or toxic agents in the air may be a contributing factor. Many studies have shown a relationship between the incidence of emphysema and bronchitis (inflammation of the bronchi) and smoking. Among British physicians death rates from bronchitis were six times higher in those smoking 25 cigarettes a day than in nonsmokers.

Effects of smoking

Measurements of lung function are significantly lower in cigarette smokers than in nonsmokers of the same age. Values for cigarette smokers are, on the average, about equal to those of nonsmokers who are 10–15 years older. There is evidence, however, that when cigarette smokers quit smoking, measurements of pulmonary function closely approach those of nonsmokers within one to two years, even in the case of heavy smokers 50–60 years old.

Kidney. The kidney removes wastes from the body by separating them from the blood and forming urine. In this process many substances are accumulated in the urine at a higher concentration than in the blood. With advancing age the concentrating ability of the kidney falls, so that a greater volume of water is required to excrete the same amount of waste material. This loss in concentrating ability is probably partially offset by a decrease in the excretory load because of reduced activity, alterations in food intake, and the reduction in muscle mass of the elderly. These changes in kidney function may not be reflected in urine volume, since volumes fluctuate widely at all ages and are determined primarily by fluid intake.

The reduction in renal (kidney) function is due in part to a gradual reduction in blood flow to the kidney. Since the kidney receives a great excess of blood (about 25 percent of the blood pumped by the heart each minute), the reduction with age does not normally result in an accumulation of waste products in the blood. Any such accumulation is the result of disease that damages the kidney. The reduced concentrating ability of the kidney results from a loss of some of the nephrons, the functional elements of the kidney, and the reduced activity of cellular enzymes.

Regulatory mechanisms. Some physiological characteristics, such as the mechanisms that regulate the acidity of the blood or its sugar level, are adequate to maintain normal levels under resting conditions even in very old people; however, the aged require more time than the young to reestablish normal levels when changes from the normal occur.

In order to test the effectiveness of control mechanisms of the body, physiologists produce changes experimentally and determine the rate of recovery. When the acidity of the blood is increased to the same extent in old and young subjects, it is returned to normal within 6–8 hours in the young; in the elderly 18–24 hours are required.

Similarly, the rate of return to fasting levels after sugar has been administered intravenously or orally is slower in the old than in the young. The response to insulin, which accelerates the removal of sugar from the blood, is also diminished in the elderly.

The body's physiological mechanisms for adjusting to changes in environmental temperature are less adequate in the old than in the young. Consequently older people may prefer more uniform and slightly higher temperatures than the young. High temperatures are also more hazardous to the elderly. The incidence of heat prostration in hot weather increases with age.

Exercise and work

Exercise is one of the physiological stresses of daily living. In reasonable amounts it is a valuable stimulus to maintain physiological vigour. A number of studies have indicated a lower incidence of cardiovascular disease among adults who indulge in physical activity than in those who do not.

The capacity to perform muscular work diminishes progressively in the elderly. Muscle strength diminishes; however, the reduction in strength is less in muscles that continue to be used throughout adult life than in those that are not. Thus a part of the reduction in muscle strength may be an atrophy of disuse.

Maximum work capacity is reduced in the elderly, largely because of the inability to deliver enough oxygen to the working muscles. In the young, the need for oxygen is met for the most part by increasing the heart rate. Under conditions of maximum work, young adults can increase their heart rate to over 200 beats per minute; the elderly to only about 150 per minute. In addition, the transfer of oxygen from the lungs to the blood is reduced in the elderly under conditions of strenuous exercise.

With less than maximum exercise, there is a greater increase in blood pressure, heart rate, and respiration in the old than in the young; that is, a given work load induces a greater physiological stress in the old than in the young. Furthermore, recovery of blood pressure, heart rate, and respiration to resting values takes longer in the old.

Premature aging. Progeria is an extremely rare disease of early childhood characterized by many of the superficial aspects of aging, such as baldness, thinning of the skin, prominence of blood vessels of the scalp, and vascular disease. These children have the general appearance of "little old men." They rarely live beyond the age of 15–18. Death is usually caused by cardiovascular disease.

The disease is extremely rare. In fact, only about 50 cases have been identified for study. In spite of the appearance of premature aging, these patients fail to show an acceleration of other age changes, except for the early development of cardiovascular disease. Most tests of physiological and psychological functions give values which are normal for their chronological age. It is doubtful whether the child with progeria is suffering only from accelerated aging; rather, progeria should be regarded as a rare disease with a superficial resemblance to senescence.

PSYCHOLOGICAL ASPECTS OF AGING

Intellectual functions in the aged

The most outstanding psychological features of aging are the impairment in short-term memory and the lengthening of response time. Both of these factors contribute to lower scores of the elderly on standard tests of "intelligence." When the aged are given all the time that they wish on tests that are not heavily dependent on school skills, their performance is only slightly poorer than that of young adults. Age decrements are negligible on tests that depend on vocabulary, general information, and well-practiced activities.

Experimental studies on learning show that, although the elderly learn more slowly than the young, they can acquire new material and can remember it as well as the young. Age differences in learning increase with the difficulty of the material to be learned.

Aged people tend to become more cautious and rigid in their behaviour and to withdraw from social contacts. These behaviour patterns may be the result of social

institutions and expectancies rather than an intrinsic phenomenon of aging. Many persons who "age successfully" make conscious efforts to maintain mental alertness by continued learning and by expansion of social contacts with individuals in a younger age group. (N.W.S.)

BIBLIOGRAPHY

Growth. L.B. AREY, *Developmental Anatomy,* 7th ed. (1965), a standard and authoritative treatment of the growth and development of the human embryo; B.I. BALINSKY, *An Introduction to Embryology,* 3rd ed. (1970), a definitive modern textbook of growth and development of animals; R.J. GOSS, *Adaptive Growth* (1964), a modern treatment of the mechanisms of compensatory growth in animals; JULIAN HUXLEY, *Problems of Relative Growth* (1932), a successor to D'Arcy Thompson's book, with formulation of the problems of allometric growth; D'ARCY WENTWORTH THOMPSON, *On Growth and Form,* abridged edition ed. by J.T. BONNER (1961), the classic mathematical treatment of the dynamics of growth, reissued in an abridged form; J.G. TORREY, *Development in Flowering Plants* (1967), a modern authoritative treatment (in paperback) of the occurrence and mechanisms of growth in higher plants.

Biological malformation. *Animal malformations:* D. BERGSMAN (ed.), "Birth Defects Original Article Series" (1965–), authoritative articles on congenital malformations and abnormalities; L.C. DUNN, "Genetic Monsters," *Scient. Am.,* 182:16–19 (1950), an illustrated account of the earliest appearance and development during embryonic life of specific abnormalities due to effects of mutated genes in vertebrates; C. STERN, *Principles of Human Genetics,* 3rd ed. (1971), illustrated descriptions of congenital abnormalities resulting from effects of mutant genes and of chromosomal aberrations; E. WOLFF, *La Science des monstres* (1948), a readable classical essay, including results of experimental teratology (in French).

Plant malformations: A.C. BRAUN, "Growth Is Affected," in J.G. HORSFALL and A.E. DIMOND (eds.), *Plant Pathology: An Advanced Treatise,* vol. 1, *The Diseased Plant,* ch. 6 (1959), a fairly comprehensive but largely descriptive account of the various types of growth abnormalities found in higher plant species; A.C. BRAUN and R.B. PRINGLE, "Pathogen Factors in the Physiology of Disease—Toxins and Other Metabolites," in C.S. HOLTON *et al.* (eds.), *Plant Pathology: Problems and Progress, 1908–1958,* ch. 9 (1959), an attempt to describe in chemical terms the specific substances involved in producing various types of abnormal growth responses in plants.

Biological regeneration. R.J. GOSS, *Principles of Regeneration* (1969), a comprehensive analysis of comparative aspects of regeneration; E.D. HAY, *Regeneration* (1966), a short and incisive account of regeneration in invertebrates and vertebrates; V. KIORTSIS and H.A.L. TRAMPUSCH (eds.), *Regeneration in Animals and Related Problems* (1965), numerous reports by outstanding authorities on all kinds of regeneration; R.M.H. MCMINN, *Tissue Repair* (1969), a thorough and up-to-date review of how various tissues of mammalian bodies heal injuries; A.E. NEEDHAM, *Regeneration and Wound-Healing* (1952), a detailed review, with numerous references; A.J. SCHMIDT, *Cellular Biology of Vertebrate Regeneration and Repair* (1968), on the cellular and molecular aspects of amphibian limb regeneration.

Biological development. Classical works that laid the foundation for the modern interpretation of development in terms of gene activities include: T.H. MORGAN, *Embryology and Genetics* (1934); R. GOLDSCHMIDT, *Physiologische Theorie der Vererbung* (1927); Eng. trans., *Physiological Genetics,* (1938); C.H. WADDINGTON, *Organisers and Genes* (1940); D.W. THOMPSON, *On Growth and Form,* rev. ed. (1942). A few more recent books, which include the theoretical as well as empirical aspects of development are: C.W. WARDLAW, *Organization and Evolution in Plants* (1965); J.T. BONNER, *Morphogenesis* (1952); E.W. SINNOTT, *The Problem of Organic Form* (1963). Probably the best general account of modern experimental work is A. KUHN, *Entwicklungsphysiologie,* 2nd ed. (1965). Less up-to-date is C.H. WADDINGTON, *Principles of Embryology* (1956). Shorter and less comprehensive works are: J.D. EBERT, *Interacting Systems in Development* (1965); and C.H. WADDINGTON, *New Patterns in Genetics and Development* (1962). A number of essays on the theory of development may be found in *Towards a Theoretical Biology,* 3 vol., ed. by C.H. WADDINGTON (1968–70), see particularly the essays by Goodwin, Thom, Wolpert, and Waddington.

Plant development. Developmental topics were taken up in many older botanical works, and special mention may be made of KARL GOEBEL, *Organographie der Pflanzen,* 2 parts (1898–1901; Eng. trans., *Organography of Plants,* 1900–05). Several modern texts on plant anatomy and morphology deal descriptively with development, including A.S. FOSTER and E.M. GIFFORD, *Comparative Morphology of Vascular Plants* (1959); K. ESAU, *Anatomy of Seed Plants* (1960), and *Plant Anatomy,* 2nd ed. (1965); and ABRAHAM FAHN, *Plant Anatomy* (1967; originally published in Hebrew, 1962). For cryptogams, general surveys of development appear in G.M. SMITH, *Cryptogamic Botany,* 2nd ed., vol. 1, *Algae and Fungi,* and vol. 2, *Bryophytes and Pteridophytes* (1955).

Morphogenesis in plant development is treated in E.W. SINNOTT, *Plant Morphogenesis* (1960); and C.W. WARDLAW, *Morphogenesis in Plants* (1968). Many recent books are devoted to physiological and biochemical aspects. Introductory treatments are provided by A.C. LEOPOLD, *Plant Growth and Development* (1964); J.G. TORREY, *Development in Flowering Plants* (1967); and P.F. WAREING and I.D.J. PHILLIPS, *The Control of Growth and Differentiation in Plants* (1970). More advanced coverage of certain parts of the subject may be found in F.C. STEWARD, *Growth and Organization in Plants* (1968); W.M. LAETSCH and R.E. CLELAND (eds.), *Papers on Plant Growth and Development* (1967); and M.B. WILKINS (ed.), *The Physiology of Plant Growth and Development* (1969).

Many important topics in plant development are best approached through specialized monographs. The following are recommended: PANCHANAN MAHESHWARI, *An Introduction to the Embryology of Angiosperms* (1950); L.J. AUDUS, *Plant Growth Substances,* 2nd ed. (1959); F.A.L. CLOWES, *Apical Meristems* (1961); and W.S. HILLMAN, *The Physiology of Flowering* (1962). Sections are devoted to development and differentiation in each volume of the *Annual Review of Plant Physiology* (1950–). The articles are advanced and require background knowledge for proper appreciation.

Animal development. General textbooks covering the whole subject of animal development include: B.I. BALINSKY, *An Introduction to Embryology,* 3rd ed. (1970); C.W. BODEMER, *Modern Embryology* (1968); and F.G. GILCHRIST, *A Survey of Embryology* (1968).

Theoretical considerations of embryonic development are found in: H. SPEMANN, *Experimentelle Beiträge zu einer Theorie der Entwicklung* (1936; Eng. trans., *Embryonic Development and Induction,* 1938, reprinted 1962), a classic exposition of the experimental method in embryology; J.T. BONNER, *The Evolution of Development* (1958); J. BONNER, *The Molecular Biology of Development* (1965); and C.H. WADDINGTON, *Principles of Embryology* (1956).

Descriptions of the morphological aspects of embryonic development are found in: L.B. AREY, *Developmental Anatomy: A Textbook and Laboratory Manual of Embryology,* 7th ed. (1965); and B.M. PATTEN, *Foundations of Embryology,* 2nd ed. (1964).

Physiological and analytical studies of embryonic development are treated in: L.J. BARTH, *Development: Selected Topics* (1964); E.M. DEUCHAR, *Biochemical Aspects of Amphibian Development* (1966); R.A. FLICKINGER (comp.), *Developmental Biology* (1966); J. NEEDHAM, *Biochemistry and Morphogenesis* (1942); L. SAXEN and S. TOIVONEN, *Primary Embryonic Induction* (1962); and J.P. TRINKAUS, *Cells into Organs: The Forces That Shape the Embryo* (1969).

Specific treatments of embryonic adaptations in a wide range of animals are considered in: W.J. HAMILTON, J.D. BOYD, and H.W. MOSSMAN, *Human Embryology,* 3rd ed. (1962), deals with the prenatal development of form and function in man; O.E. NELSEN, *Comparative Embryology of the Vertebrates* (1953); and M. KUME and K. DAN (eds.), *Invertebrate Embryology* (1968; originally published in Japanese, 1957).

Books dealing with metamorphosis and senescence include the following: W. ETKIN, "Metamorphosis," in J.A. MOORE (ed.), *Physiology of the Amphibia,* ch. 8 (1964); V.B. WIGGLESWORTH, *The Physiology of Insect Metamorphosis* (1954); and A. COMFORT, *Ageing: The Biology of Senescence,* rev. ed. (1964).

Human embryology. L.B. AREY, *Developmental Anatomy,* 7th ed. (1965), a comprehensive, modern text on human development; B.I. BALINSKY, *An Introduction to Embryology,* 3rd ed. (1970), a college-level treatment of development with emphasis on early stages and comparative conditions; E.B. BLECHSCHMIDT, *The Stages of Human Development Before Birth* (1961), an atlas of human embryos, sections, and reconstructed organ systems; G.W. CORNER, *Ourselves Unborn* (1944), a popular description of human development, its biological significance, and its aberrations; W.J. HAMILTON *et al., Human Embryology,* 4th ed. (1972), a well-illustrated account of human development; J.E. MORISON, *Foetal and Neonatal Pathology,* 2nd ed. (1963), a discussion of the disturbances of prenatal life and of adaptations to extrauterine existence; J. NEEDHAM, *A History of Embryology,* 2nd ed. rev. (1959), a historical review from antiquity through the 18th century; B.M. PATTEN, *Human Embryology,* 3rd ed. (1968), a largely structural account of human development with emphasis on the nervous and cardiovascular systems; A. SORSBY (ed.), *Clinical Genetics* (1953), a compilation of known relations of heredity to disease and abnormal development; J.G. WILSON, "Experimental Studies on Congenital Malformation," *J. Chron. Dis.,* 10:110–130 (1959),

a summary of recent discoveries on causative factors related to abnormal development.

Human development. B.T. DONOVAN and J.J. VAN DER WERF TEN BOSCH, *Physiology of Puberty* (1965), a standard monograph dealing with events and control of puberty for the most part in experimental animals; F. FALKNER (ed.), "Child Development: An International Method of Study," *Annales Paediatrici,* suppl. no. 72 (1960), a description with illustrations of the methods of taking body measurements of children; *Human Development* (1966), a text with chapters by many contributors varying somewhat in their approach and covering both psychological and physical development; W.A. MARSHALL and J.M. TANNER, "Variations in Pattern of Pubertal Changes in Girls," *Arch. Dis. Child.,* 44:291–303 (1969), and "Variations in the Pattern of Pubertal Changes in Boys," *Arch. Dis. Child.,* 45:13–23 (1970), descriptive papers; J.M. TANNER, *Education and Physical Growth* (1961), a standard textbook on physical growth, written for teachers and used for colleges of education but on a level also appropriate for the layman; *Growth at Adolescence,* 2nd ed. (1962), a more advanced textbook than the above, suitable chiefly for pediatricians and biologists; "Earlier Maturation in Man," *Scient. Am.,* 218:21–27 (1968), a popular article on earlier maturing; "Growth and Endocrinology of the Adolescent," in L. GARDNER (ed.), *Endocrine and Genetic Diseases of Childhood* (1969), an up-to-date and advanced description in a standard medical textbook; J.M. TANNER and G.R. TAYLOR, *Growth* (1965), readings for high school students and laymen distinguished by beautiful illustrations; J.M. TANNER, R.H. WHITEHOUSE, and M. TAKAISHI, "Standards from Birth to Maturity for Height, Weight, Height Velocity and Weight Velocity, British Children 1965," *Arch. Dis. Child.,* 41:454–471 (1966), an advanced and somewhat difficult paper which forms the basis for most of the standards used in present day pediatrics.

Life-span. *Human beings:* WARREN ANDREWS, *Anatomy of Aging in Man and Animals* (1970); SEYMOUR BAKERMAN (ed.), *Aging Life Processes* (1969); AUSTIN M. BRUES and G.A. SACHER (eds.), *Aging and Levels of Biological Organization* (1965); ALEXANDER COMFORT, *The Biology of Senescence* (1956); L.I. DUBLIN, A.J. LOTKA, and M. SPIEGELMAN, *Length of Life,* rev. ed. (1949); E.S. DEEVEY, JR., "The Probability of Death," *Scient. Am.,* 182:58–60 (1950); NATHAN W. SHOCK (ed.), *Biological Aspects of Aging,* 2nd ed. (1962).

Subhuman animals: W.C. ALLEE *et al., Principles of Animal Ecology* (1949); F.S. BODENHEIMER, *Problems of Animal Ecology* (1938); CIBA FOUNDATION, *The Lifespan of Animals* (1959); ALEX COMFORT, "The Life Span of Animals," *Scient. Am.,* 205:108–119 (1961); E.S. DEEVEY, JR., "Life Tables for Natural Populations of Animals," *Q. Rev. Biol.,* 22:283–314 (1947); D.L. LACK, *The Natural Regulation of Animal Numbers* (1954).

Plants: A.E. PORSILD, C.R. HARINGTON, and G.A. MULLIGAN, "Lupinus arcticus Wats. Grown from Seeds of Pleistocene Age," *Science,* 158:113–114 (1967); W.J. ROBBINS, "Physiological Aspects of Aging in Plants," in W.C. STEERE (ed.), *Fifty Years of Botany* (1958); B.L. STREHLER *et al.* (eds.), *Biology of Aging* (1960); FRITZ W. WENT, "Plant Growth and Plant Hormones," in W.H. JOHNSON and W.C. STEERE (eds.), *This Is Life* (1962); A. CARL LEOPOLD *et.al., Plant Growth and Development,* 2nd ed. (1975).

Aging. J.A. BEHNKE, C.B. FINCH and G.B. MOMENT (eds.), *The Biology of Aging* (1978), a collection of articles covering many aspects of the subject, including aging in cells and molecules, in plants and animals, and evolutionary considerations; ALEXANDER COMFORT, *The Biology of Senescence,* 3rd ed. (1978), the best general treatise, except for biochemical mechanisms; R.R. KOHN, *Principles of Mammalian Aging,* 2nd ed. (1978), coverage of the fundamentals with emphasis on the role of interstitial tissue changes in the aging process; M.J. LAMB, *Biology of Aging* (1977), a concise introduction to the mechanisms of aging; P.J. LINDOP and G.A. SACHER (eds.), *Radiation and Ageing* (1966), pathological changes with aging, life table characteristics of laboratory animal populations, and the influence of ionizing radiations; NATHAN W. SHOCK (ed.), *Perspectives in Experimental Gerontology* (1966), a volume of collected brief essays on aging; B.L. STREHLER, *Time, Cells, and Aging,* 2nd ed. (1977), an examination of cellular and molecular mechanisms and theories of aging; G.E.W. WOLSTENHOLME and MAEVE O'CONNOR (eds.), *Ciba Foundation Colloquia on Ageing,* 5 vol. (1955–59), the best source for molecular and cellular theories and data: vol. 1 covers *General Aspects;* 2, *Aging in Transient Tissues;* 3, *Methodology;* 4, *Water and Electrolyte Metabolism;* and 5, *Lifespan of Animals,* with some consideration of ecological and evolutionary aspects; *United Nations Demographic Yearbook* includes mortality statistics for all countries, showing the influence of economic, social, and climatic factors on mortality.

Human aging. Additional information on this subject may be found in the following texts: J.E. BIRREN (ed.), *Handbooks of Aging,* 3 vol. (1976); A. COMFORT, *The Biology of Senescence,* 3rd. ed. (1978); H.J. CURTIS, *Biological Mechanisms of Aging* (1966); A.M. HOFFMAN (ed.), *The Daily Needs and Interests of Older People* (1970); E.V. COWDRY, *Problems of Ageing: Biological and Medical Aspects* (1939, reprinted 1979; 3rd ed. by A.I. LANSING, 1952); R.W. PREHODA, *Extended Youth: The Promise of Gerontology* (1968); M.W. RILEY (ed.), *Aging from Birth to Death* (1979), and, with A. FONER, *Aging and Society,* 3 vol. (1968–72); N.W. SHOCK, *Trends in Gerontology,* 2nd ed. (1957, reprinted 1976); B.L. STREHLER, *Time, Cells, and Aging,* 2nd ed. (1977); and P.S. TIMIRAS, *Developmental Physiology and Aging* (1972).

Guyana

Guyana (the Co-operative Republic of Guyana) is an independent republic and member of the Commonwealth located in the northeastern corner of South America. It is bordered by Venezuela to the west, Brazil to the southwest and south, Suriname (along the Courantyne River) to the east, and the Atlantic Ocean to the north. Its total area of 83,000 square miles (215,000 square kilometres) is largely uninhabited, and most of the country's inhabitants occupy the narrow coastal strip. The capital and chief port is Georgetown.

Present-day Guyana reflects its British colonial past and its reactions to that past. It is the only English-speaking country of South America. Since independence in 1966, Guyana's chief economic assets—its sugarcane plantations and bauxite industry—have come under government control, as has most of the country's commerce. Guyana's populace is mainly of colonial origin, although a small number of aboriginal Indians are scattered throughout the forested interior.

The more numerous coastal peoples are chiefly descendants of slaves from Africa and indentured workers from India, who were originally imported to work the coastal sugarcane plantations. Racial problems between the latter two groups have played a disruptive role in Guyanaese society.

Politically, Guyana has moved on a steady course toward socialism from the time of independence, although after the death of the first prime minister, Forbes Burnham, in 1985, ties with Western powers were strengthened.

This article is divided into the following sections:

Physical and human geography 441
- The land 441
  - Relief
  - Drainage
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 443
- The economy 443
  - Resources
  - Agriculture, forestry, and fishing
  - Industry
  - Finance and trade
  - Transportation
- Government and social conditions 445
  - Government
  - Education
  - Health and welfare
- Cultural life 445

History 445

Bibliography 446

Physical and human geography

THE LAND

Relief. The narrow plain that extends along the country's Atlantic coast has been modified considerably by humans. Much of the area, which measures only about 10 miles (16 kilometres) at its widest point, has been reclaimed from the sea by a series of canals and some 140 miles of dikes. The coastal plain's inland border is generally marked by canals that separate the plain from interior swamps. South of the coastal zone the forested land rises gently and has sandy soils.

The coastal regions

About 40 miles inland from the coast is a region of undulating land that rises from 50-foot (15-metre) hills on the coastal side of the region to 400-foot (120-metre) ones on the western side. The area is between 80 and 100 miles wide and is widest in the southeast. It is covered with sands, from which it takes its name as the white-sands (*zanderij*) region. A small savanna region in the east lies about 60 miles from the coast and is surrounded by the white-sands belt. The sands partly overlie a low crystalline plateau that is generally less than 500 feet in elevation. The plateau forms most of the country's centre and is penetrated by igneous rock intrusions that cause the numerous rapids of Guyana's rivers.

Beyond the crystalline plateau, the Kaieteurian Plateau lies generally below 1,600 feet above sea level; it is the site of the spectacular Kaieteur Falls, noted for their sheer 741-foot initial plunge. The plateau is overlain with sandstones and shales that, in the south, form the extensive Rupununi Savanna region. The Acarai Mountains (Serra Acaraí), which rise to about 2,000 feet, rim the plateau on the southern border, and it is crowned on the western frontier by the Pakaraima Mountains, which rise to 9,094 feet (2,772 metres) in Mount Roraima. The Rupununi Savanna is bisected by the east–west Kanuku Mountains, which rise to almost 3,000 feet.

Major rivers

Drainage. Guyana's four main rivers—the Courantyne, Berbice, Demerara, and Essequibo—all flow from the south and empty into the Atlantic along the eastern section of the coast. Among the tributaries of the Essequibo, the Potaro, Mazaruni, and Cuyuni drain the northwest, and the Rupununi drains the southern savanna. The coast is cut by shorter rivers, including the Pomeroon, Mahaica, Mahaicony, and Abary.

The rivers are part of the watershed of the Amazon and Orinoco rivers, and the headwaters of the Rupununi in Brazil are often confused with those of the Amazon. Drainage is poor, because the average gradient is only one foot per mile, and there are swamps and flooding in the mountains and savannas. The rivers are not suitable for long-distance transportation because they are broken by interior falls, and in the coastal zone their mouths and estuaries are blocked by mud and by sandbars that may occur two to three miles out to sea.

Soils. The coastal soils are fertile but acidic. The fine-particle, grayish blue clays of the coastal plain are composed of alluvium from the Amazon deposited by the south equatorial ocean current and of much smaller amounts of alluvium from the country's rivers. They overlie white sands and clays and can support intensive agriculture but must be subjected to fallowing to restore fertility. *Pegasse* soil, a type of tropical peat, occurs behind the coastal clays and along the river estuaries, while silts line the banks of the lower rivers. Reef sands occur in bands in the coastal plain, especially near the Courantyne and Essequibo rivers. The rock soils of the interior are leached and infertile, and the white sands are almost pure quartz.

Coastal soils

Climate. High temperatures, heavy rainfall with small seasonal differences, high humidity, and high average cloud cover provide climatic characteristics of an equatorial lowland. Temperatures are remarkably uniform. At Georgetown the daily temperature varies from 74° to 86° F (23° to 30° C), and the mean temperature is about 80° F (27° C). The constant heat and high humidity are mitigated near the coast by the trade winds.

Rainfall derives mainly from the movement of the intertropical front, or doldrums. It is heavy everywhere on the plateau and the coast. The annual average at Georgetown is about 90 inches (2,290 millimetres), and on the interior Rupununi Savanna it is about 70 inches. On the coast a long wet season, from April to August, and a short

wet season, from December to early February, are sufficiently well marked on the average, but in the southern savannas the short wet season does not occur. Total annual rainfall is variable, and seasonal drought can occur in July and August when the southeast trade winds parallel the coast. Variations in Guyana's climatic patterns have a determining effect on tropical crop production.

Plant and animal life. Many plants of the coast, such as the mangrove and various saltwater grasses, grow in shallow brackish water and help to protect or extend the land. The wet savanna behind the coast has coarse tufted grasses and a wide scattering of palms, notably the coconut, truli, and manicole. High rain forest, or selva, covers about three-fourths of the land area and is of extraordinary variety and magnificence. Prominent trees include the greenheart and the wallaba on the sandy soils of the northern edge, the giant mora and crabwood on swampy sites, the balata and other latex producers, and many species such as the siruaballi and hubaballi that yield handsome cabinet woods. The interior savanna is mostly open grassland, with much bare rock, many termite hills, and clumps of ita palm.

The rain forest

All forms of animal life are immensely varied and abundant, though few, apart from birds and insects, are normally visible. The tapir is the country's largest land mammal, and the jaguar is the largest and fiercest of the cats, which also include the ocelot; monkeys and deer are the most common animals. Among the more exotic species are the sloth; great anteater; the capybara, or bush pig; and armadillo. Birds include the vulture, kiskadee, blue sacki, hummingbird, kingfisher, and scarlet ibis of the coast and lower rivers; and the macaw, tinamou, bellbird, and cock-of-the-rock in the forest and savanna. The caiman (a reptile similar to the alligator) is the most common of the larger freshwater creatures. The giant anaconda, or water boa, is the largest of the many kinds of

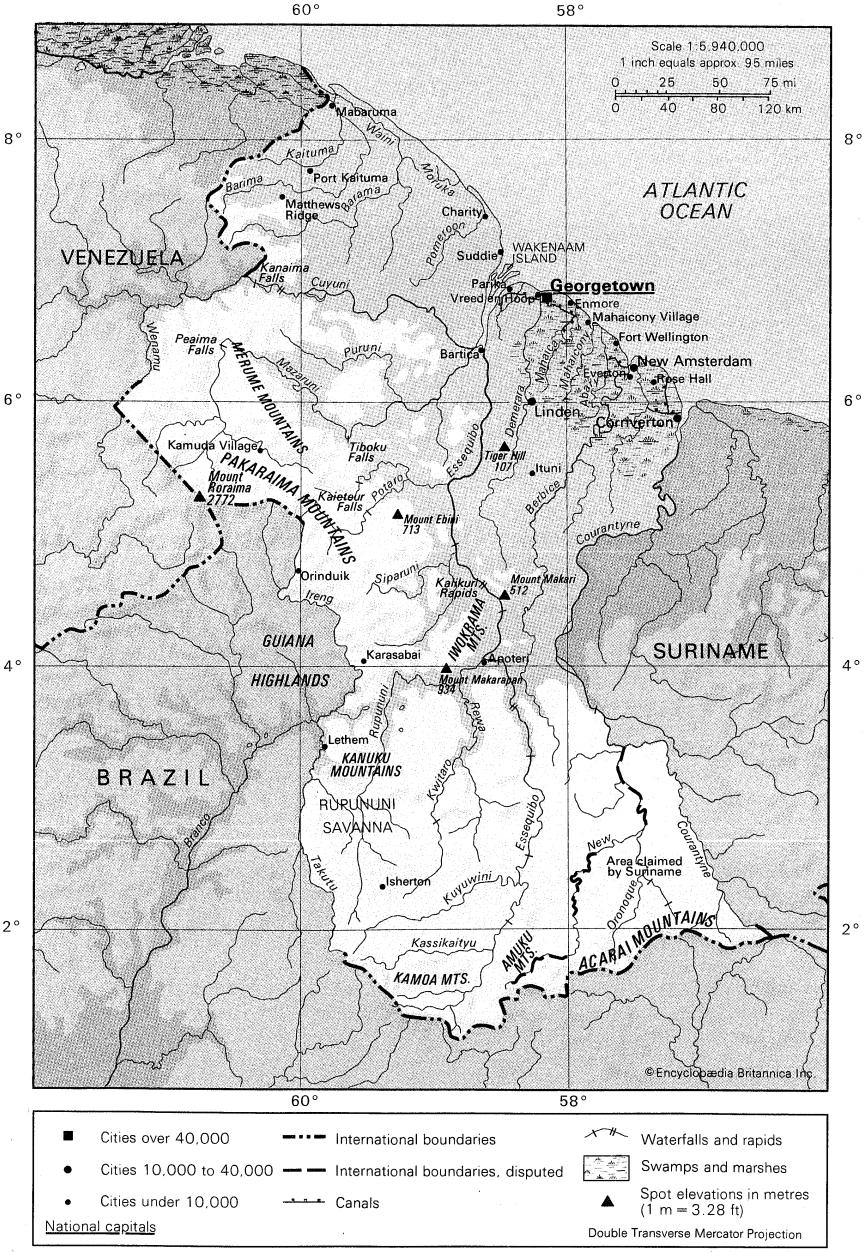
MAP INDEX

Cities and towns

Apoteri ..... 04 02 N 58 34 W
Bartica ..... 06 24 N 58 37 W
Charity ..... 07 24 N 58 36 W
Corriverton ..... 05 52 N 57 10 W
Enmore ..... 06 46 N 57 59 W
Everton ..... 06 12 N 57 31 W
Fort Wellington .. 06 24 N 57 36 W
Georgetown .... 06 48 N 58 10 W
Isherton ..... 02 19 N 59 22 W
Ituni ..... 05 30 N 58 14 W
Kamuda Village .. 05 38 N 60 18 W
Karasabai ..... 04 02 N 59 32 W
Lethem ..... 03 23 N 59 48 W
Linden ..... 06 00 N 58 18 W
Mabaruma ..... 08 12 N 59 47 W
Mahaicony Village ..... 06 36 N 57 48 W
Matthews Ridge . 07 30 N 60 10 W
New Amsterdam . 06 15 N 57 31 W
Orinduik ..... 04 42 N 60 01 W
Parika ..... 06 52 N 58 25 W
Port Kaituma .... 07 44 N 59 53 W
Rose Hall ..... 06 16 N 57 21 W
Suddie ..... 07 07 N 58 29 W
Vreed en Hoop .. 06 48 N 58 11 W

Physical features and points of interest

Abary, *river* ..... 06 33 N 57 44 W
Acarai Mountains ..... 01 50 N 57 30 W
Amuku Mountains ..... 01 45 N 58 20 W
Atlantic Ocean .. 07 30 N 57 00 W
Barama, *river* .... 07 40 N 59 15 W
Barima, *river* .... 07 30 N 60 00 W
Berbice, *river* .... 06 17 N 57 32 W
Courantyne, *river* ..... 05 57 N 57 06 W
Cuyuni, *river* .... 06 23 N 58 41 W
Demerara, *river* .. 06 48 N 58 10 W
Ebini, Mount .... 05 11 N 59 15 W
Essequibo, *river* . 06 59 N 58 23 W
Guiana Highlands ..... 04 00 N 60 00 W
Ireng, *river* ..... 03 33 N 59 51 W
Iwokrama Mountains ..... 04 20 N 58 44 W
Kaieteur Falls ... 05 10 N 59 28 W
Kaituma, *river* ... 08 11 N 59 41 W
Kalikuri Rapids .. 04 39 N 58 39 W
Kamoa Mountains ..... 01 37 N 59 00 W
Kanaima Falls ... 06 53 N 60 15 W
Kanuku Mountains ..... 03 12 N 59 30 W
Kassikaityu, *river* . 01 49 N 58 32 W
Kuyuwini, *river* ... 02 16 N 58 16 W
Kwitaro, *river* .... 03 19 N 58 47 W
Mahaica, *river* ... 06 43 N 57 55 W
Mahaicony, *river* . 06 36 N 57 48 W
Makarapan, Mount ..... 04 00 N 58 51 W
Makari, Mount ... 04 32 N 58 23 W
Mazaruni, *river* .. 06 25 N 58 38 W
Merume Mountains ..... 05 48 N 60 06 W
Moruka, *river* .... 07 40 N 58 48 W
New, *river* ..... 03 23 N 57 36 W
Oronoque, *river* .. 02 45 N 57 25 W
Pakaraima Mountains ..... 04 05 N 61 30 W
Peaima Falls .... 06 21 N 60 36 W
Pomeroon, *river* .. 07 37 N 58 45 W
Potaro, *river* .... 05 22 N 58 54 W
Puruni, *river* .... 06 00 N 59 12 W
Rewa, *river* ..... 03 53 N 58 45 W
Roraima, Mount .. 05 12 N 60 44 W
Rupununi, *river* .. 04 03 N 58 34 W
Rupununi Savanna, *region* ..... 03 00 N 59 30 W
Siparuni, *river* ... 04 50 N 58 49 W
Takutu, *river* .... 02 50 N 60 00 W
Tiboku Falls .... 05 43 N 59 38 W
Tiger Hill ..... 05 39 N 58 23 W
Waini, *river* ..... 08 24 N 59 51 W
Wakenaam Island ..... 06 58 N 58 27 W
Wenamu, *river* ... 06 43 N 61 07 W



snakes, and the bushmaster is the most vicious. Lizards are numerous and include the iguana in the lower rivers. Sharks and stingrays are found offshore. The snapper and grouper are common ocean fish, and shrimps abound in the muddy currents off the coast. The manatee is also common in Guyanaese waters. Among the freshwater fish is the huge piraucu, which attains lengths up to 14 feet.

Settlement patterns. The country is divided traditionally between the coast, where most of the population is concentrated, and the interior. The coastal population is heterogeneous, its inhabitants descended from the labourers brought in to work the sugarcane plantations. The interior, despite scattered ranching and mining settlements, is largely the province of the indigenous Indians.

Rural society

Guyanese society is predominantly rural, most of the people occupying villages in the coastal region. The highest population concentrations are along the estuary of the Demerara River and between the mouths of the Berbice and Courantyne rivers. Village units are of distinctive rectangular shapes, with the settlement areas nearest the ocean and connected to one another by the coastal highway; each village's farmlands extend inland, often for several miles, and are separated from neighbouring village lands by canals. Villages range in size from several hundred to several thousand persons. The commonly found wood and concrete-block dwellings are usually built on stilts above the flood-prone land and are connected by footbridges to the streets, which are built over the drainage and irrigation canals.

Georgetown is the country's main port and its largest city. Located at the mouth of the Demerara River, it lies below sea level and is protected by dikes along both the river and the sea. Other important towns include the interior bauxite-mining centre of Linden and the market centre of New Amsterdam, located on the mouth of the Berbice River. Agricultural centres, such as the sugarcane plantation of Port Mourant, east of New Amsterdam, and the rice centre of Anna Regina, north of the Essequibo River estuary, provide commercial and marketing functions in the rural areas of the coastal zone.

THE PEOPLE

The indigenous peoples of Guyana are collectively known as Amerindians and constitute about 4 percent of the population. Indian groups include the Warao (Warrau), Arawak, Carib, Wapisiana (Wapishana), Arecuna, the mixed "Spanish Arawak" of the Moruka River, and many more in the forest areas. The Makusí (Macussí or Macushí) are the most prominent of the savanna peoples. Sizable concentrations of Amerindians inhabit the far west along the border with Venezuela and Brazil. They are rarely seen in the populated coastal areas, although a few have interbred with blacks and East Indians. Since 1970, traditional Amerindian lands near the international borders have come under government control, although Amerindians continue to hold village lands informally throughout Guyana's interior.

Chief ethnic groups

The other major elements in the population are predominantly coastal dwellers. Descendants of African slaves form the oldest group; they abandoned the plantations after full emancipation in 1838 to become independent peasantry or town dwellers. The Afro-Guyanese constitute about one-third of the population. The East Indians came mostly as indentured labour from India to replace Africans in plantation work. They form the largest racial group in the country—about half the population—and have been increasing more rapidly than the others. The East Indians are the mainstay of plantation agriculture, and many are independent farmers and landowners, have done well in trade, and are well represented among the professions.

The Chinese and Portuguese also entered originally as agricultural labourers but are now rarely found outside the towns. They are active in business and the professions, and their influence is disproportionate to their numbers; they have not been increasing, however, and together they constitute only a tiny percentage of the population. Europeans other than Portuguese are few, and most are short-term inhabitants. While every kind of racial mixture may be found, mulattoes (persons of mixed white and black ancestry) are by far the most common. Most of them live in towns, and a high proportion are in clerical or professional work.

The major religions are Christian (chiefly Anglican and Roman Catholic) and Hindu. Fundamentalist Protestantism has made inroads in the 20th century, mainly in Georgetown. There is also a sizable minority of Muslims. Animistic religions are still practiced by some of the Amerindian peoples. The official and principal language is English, but a creole patois is spoken throughout the country. Hindi and Urdu are heard occasionally among older East Indians.

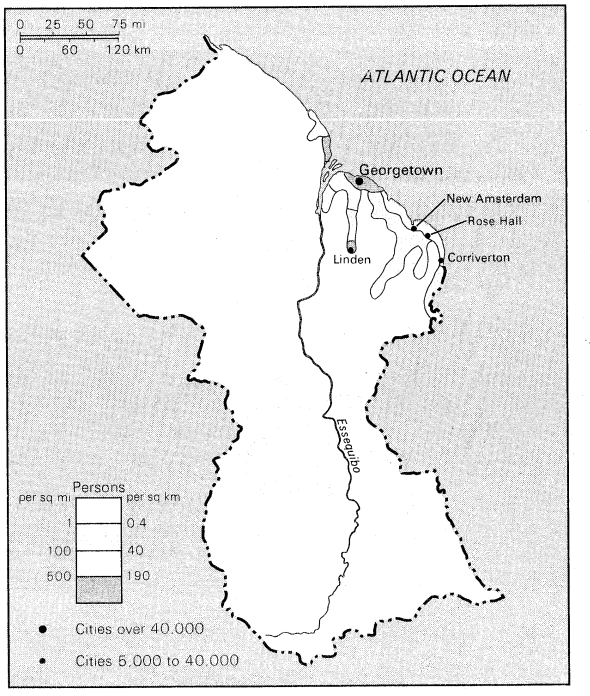
Immigration is no longer important, and by the late 20th century the number of foreign-born, long-term residents was insignificant. Emigration has been a drain on the country's human resources as thousands of persons have left annually, going mainly to the United States, Canada, England, and Caribbean islands. Many of the emigrants have been skilled and professional people whose loss has intensified the country's severe economic problems. East Indians have emigrated in large numbers to flee what they consider political persecution, a number of them having sought part-time work across the Courantyne River in western Suriname.

THE ECONOMY

Since independence Guyana has remained locked into a typical colonial economic dependency on agricultural and mined products, most notably sugarcane and bauxite. Independence brought economic reforms under a socialist-leaning government, but the effect on the old economic cycle has been minimal. Although the government permits a three-sector economy—private, public, and cooperative—the public sector remains heavily dominant.

Nationalization in the 1970s

Government management of the economy has become direct and significant. During the 1970s the government nationalized U.S. and Canadian bauxite holdings; in 1976 it nationalized the vast holdings of the Booker McConnell companies in Guyana, which included coastal sugarcane plantations as well as an array of light manufacturing and commercial enterprises. By the mid-1980s it was estimated that the government controlled directly more than 80 percent of Guyana's economy. All nationalized businesses have been reorganized under the Guyana State Corporation. The state-owned Guyana Sugar Corporation controls the sugarcane plantations, and the Guyana Mining Enterprise Ltd. was established to oversee local mineral production.



Population density of Guyana.

The Guyanese economy has deteriorated under government management policies. Members of the ruling People's National Congress (PNC) political party have been placed in managerial positions, leading to the exodus of former managers and clerical workers. Declining output, a reliance on volatile external commodity markets, and a reduced tax base have all increased financial deficits. External debt has risen precipitously, and a devalued currency has been eroded by speculation in the local black market. Reduced fuel imports have led to widespread power outages, and a government austerity program all but eliminated imported food and consumer goods. Guyana's per capita income (estimated at about $600 in the late 1980s) places it among the world's poorest countries. Improvements in economic conditions became dependent upon foreign aid and a variety of regional and reciprocal trade agreements.

Trade associations

Trade associations have an important influence in Guyanese government. The Trade Union Congress is an association of major unions, among which are the Guyana Mine Workers' Union, which is composed almost exclusively of black workers, and the Guyana Agricultural and General Workers' Union is a predominantly East Indian association.

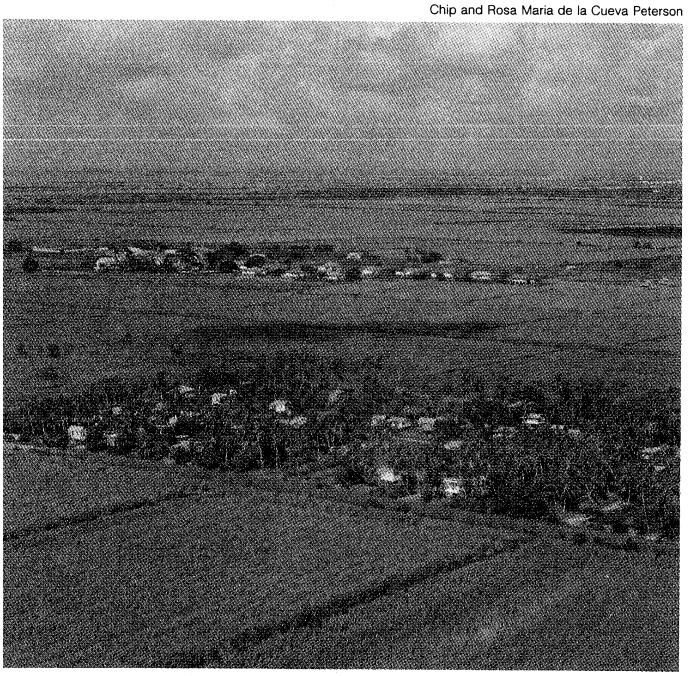
Resources. The most important mineral resource is the extensive bauxite deposits between the Demerara and Berbice rivers. There are also significant deposits of manganese at Matthews Ridge in the northwest, about 30 miles east of the Venezuelan frontier. Diamonds occur in the Mazaruni and other rivers of the Pacaraima Mountains. Gold is found in both alluvial and subsurface deposits. Other minerals include copper, iron ore, molybdenite (the source of molybdenum), nickel, white sand (used in glass manufacture), kaolin (china clay), and graphite. The government has encouraged oil exploration, but no significant reserves have been found.

The main biological resource consists of the hardwoods of the tropical rain forest and especially the greenheart tree, which is resistant to termites, decay, and marine erosion. The shrimps off the coast and a few inland fishes form the basis of the nation's fishing industry, and the grasses of the savanna regions are used for cattle grazing.

Sources of energy

Most of Guyana's energy must be imported; domestic electricity is produced largely by thermal generation and is available only on the coastal plain and along the lower reaches of the rivers. Hydroelectric potential in Guyana is considerable, especially at Tiger Hill on the Demerara River and Tiboku Falls on the Mazaruni. Development is hampered, however, by the remoteness of the falls and the large amounts of capital needed for generation and transmission facilities.

Chip and Rosa Maria de la Cueva Peterson



Sugarcane fields and communities at Ogle, east of Georgetown.

Agriculture, forestry, and fishing. Agriculture is concentrated on the narrow sea-level coastal plain between the Essequibo and Courantyne rivers. Land-use patterns still reflect early Dutch and British water-control techniques. Arable land is laid out in strips between the sea or a river and inland swamps. It is protected on all sides by dikes and canals that are used for both irrigation and drainage. The land reclaimed from the sea is fertile but acidic; lost fertility must be returned to the soil by periodic fallowing or the addition of fertilizers.

Food crops include cassava, corn (maize), bananas, vegetables, and citrus fruits. Cash crops are mainly sugarcane and rice but also include coffee and cacao. Both sugarcane and rice are cultivated through a combination of mechanization and hand labour. Agricultural production increased during the mid-20th century, mainly because mechanization extended cultivable lands, although output stagnated in later decades as the entire economy foundered. East Indian workers overwhelmingly predominate in agriculture.

Livestock production is carried out on the Rupununi Savanna and on the coastal plain. Animals include beef cattle, dairy cattle, pigs, goats, sheep, and poultry.

Forestry activities are hampered by the lack of adequate transportation, the difficulty of cutting the extremely hard wood of Guyana's trees, and the shortage of facilities for the sawing, storing, and shipping of timber. Most of the timber produced for the domestic market and for export is from the greenheart tree.

Many fishing facilities have been improved, and total production has increased as fishing has become a more important part of the economy. Shrimping is carried out primarily for export.

Industry. Guyana is one of the world's largest producers of bauxite. All alumina (aluminum oxide occurring in hydrated form in bauxite) and most of the bauxite mined is produced at Linden. The rest of the country's bauxite mining takes place on the Berbice River; a processing plant also operates downriver at Everton.

Bauxite production

Diamonds continue to be mined by hand and by suction dredges in the interior rivers. Gold is mined by individual prospectors, and large-scale Canadian-financed gold mines were opened late in the 1980s.

The country's many rice mills, like its rice fields, are generally small-scale and individually owned, although there are several large government mills along the coast. Other domestic industries are oriented toward the replacement of consumer imports such as cigarettes and matches, edible oils, margarine, beverages, soap and detergents, and clothing. Refined sugar, stock feeds, and rum and beer are also produced.

Finance and trade. The Bank of Guyana, established in 1965, has the sole right of note issue and acts as banker to the government and other banks. The country's major commercial banks include three local banks and branches of Canadian and Indian banks. Other financial services are provided by the Guyana Cooperative Agricultural and Industrial Development Bank and the New Building Society; insurance companies, most of which are foreign-owned; and more than 1,500 cooperative societies, which serve as savings institutions and offer agricultural credit.

Guyana's major trading partners are the United States, the United Kingdom, and Trinidad and Tobago. Guyana joined the Caribbean Free Trade Association (Carifta) in 1965 and then became a member of the Caribbean Community and Common Market (Caricom), which replaced Carifta in 1973. The major exports are bauxite and alumina, sugar, and rice. Shrimps, diamonds, molasses, rum, and timber are also sold abroad. Major imports include fuels and lubricants, machinery, vehicles, textiles, and foods.

Transportation. The limited road and highway system is partly paved and partly made of burnt clay. The few hundred miles of paved roads are mostly in the coastal zone. The interior has few roads.

Guyana's coastal railway, established in 1848 as South America's first rail line, was discontinued in the 1970s,

ending passenger service. A remaining freight line connects the manganese mines at Matthews Ridge with Port Kaituma on the Kaituma River, and another transports bauxite between Ituni and Linden.

Air and marine transport

Guyana Airways Corporation operates scheduled domestic and international flights. Timehri International Airport, established in 1968 and located 25 miles from Georgetown, is the country's main airport and is served by several international airlines. Domestic commercial and private aircraft, chiefly carrying passengers and equipment, use landing strips and the quieter stretches of rivers.

Barges and small boats carry people and agricultural products in the canals of the coastal estates and villages. Larger boats traverse the estuaries that intersect the coastal plain. A pontoon bridge across the Demerara River opened in 1978; it is the only bridge to link major segments of the coastal plain. Bauxite is loaded into oceangoing ships at Linden and manganese ore at Port Kaituma, but otherwise the country's external trade passes through Georgetown, which maintains connections with the West Indies, Suriname, French Guiana, the United Kingdom, Canada, and the United States.

GOVERNMENT AND SOCIAL CONDITIONS

Government. Guyana became an independent member of the Commonwealth in 1966 and in 1970 became a cooperative republic, involving citizens' organizations in government. Under the constitution of Oct. 6, 1980, executive power is vested in the president, who leads the majority party in the unicameral National Assembly and holds office for the assembly's duration. The president appoints the Cabinet, which is responsible to the National Assembly. The minority members of the assembly elect an opposition leader. The assembly is elected by universal adult suffrage for a term of five years.

The right to vote belongs to all Guyanese citizens 18 years of age or older. Voting is carried out by secret ballot under a system of proportional representation. Votes are cast for lists of candidates compiled by the political parties, and seats are allocated proportionally among the lists.

Political parties

Since independence in 1966, Guyana has been ruled by one party, the People's National Congress. Initially identified with the urban black populace, the PNC essentially established a one-party state under the direction of its first leader, Forbes Burnham. The PNC won power in an election marked by numerous reports of irregularities, many of which were related to the Guyana Defence Force (GDF), a military unit established in 1965 with strong ties to the PNC. Both the GDF and the police force are overwhelmingly black.

The People's Progressive Party (PPP), the PNC's official opposition, is the traditional party of the rural East Indians; smaller parties include the Working People's Alliance (WPA), a newer party founded by the historian Walter Rodney and headed by black labour leaders and intelligentsia allied against alleged PNC corruption.

Local government is administered principally through the Regional Democratic Councils, each led by a chairman; they are elected for terms of up to five years and four months in each of the country's 10 regions.

Guyana has two legal traditions, the British common law and the Roman-Dutch code, the latter now largely relegated to matters of land tenure. The constitution is the supreme law of the land. The court structure consists of magistrate courts for civil claims of small monetary value and minor offenses, the High Court, with original and appellate jurisdiction in civil and criminal matters, and the Court of Appeal, with appellate authority in criminal cases. The Court of Appeal and the High Court together constitute the Supreme Court.

Education. Education is free and compulsory. Primary and secondary instruction are separate, although the lack of facilities makes it necessary to hold some secondary classes in primary schools. In 1976 the government assumed full responsibility for education from nursery school to university. Government authority was extended over all church and private primary schools. Teachers are expected to teach loyalty to both the PNC and socialist objectives. The principal university is the University of Guyana, founded in 1963 and subsequently housed at Turkeyen, in the eastern part of Greater Georgetown. The school has also become politicized, attendance there being contingent upon prospective students completing a year of national service, usually at camps in Guyana's interior. Thus many Guyanese seek education and training abroad. There are also a number of other colleges, including technical and teacher-training schools.

Government control of education

Health and welfare. Health standards declined after independence. Many doctors and other trained personnel have emigrated, and economic austerity programs have reduced supplies of medicine and soap. Food shortages have created widespread malnutrition, especially in Georgetown. Diseases formerly under control, notably beriberi and malaria, had reappeared by the early 1980s, and sanitation problems have also increased.

Under colonial rule public health was centred around government and plantation health clinics. After independence a universal health care system was instituted, and most hospital facilities came under government control. Health problems arise particularly along the easily flooded coast, where the many ditches and ponds provide ideal environments for the spread of disease. A minimal government pension plan for the sick and aged has continued beyond independence, its effectiveness reduced by inflation. Government housing projects, confined mainly to the Georgetown area, have not produced expected results.

The national social structure was inherited from the period of British colonial rule, under which the majority of East Indian and Afro-Guyanese labourers were directed by white planters and government officials. A poorly defined local middle class composed of teachers, professionals, and civil servants, and including a disproportionate number of Chinese and Portuguese, emerged during colonialism. Since independence the PNC political elite has replaced the white plantocracy at the apex of Guyana's social order. The Amerindians remain apart from the country's social structure as they did under the British.

CULTURAL LIFE

Colonial influence

Postindependence Guyanese culture still bears the imprint of its colonial heritage. Guyanese were taught to respect and covet European values during the colonial era, and this has not changed despite government exhortation. Yet ethnic identity continues to be important, with daily life centring around ethnic and family groups; the mother- and grandmother-dominated family among blacks differs from the father-oriented East Indian family. Men of both groups often commute long distances to work along the coastal highway. Daily dress normally does not distinguish one group from another.

Amerindian culture, which remains uninfluenced by national politics, is recognized as an important element in Guyanese museum displays and as an inspiration in local music and painting. Cultural institutions are concentrated in Georgetown, including the Guyana Museum, which includes the Guyana Zoo, with its impressive collection of animals from northern South America. Guyanese writers have made noteworthy contributions to literature; the works of Wilson Harris, A.J. Seymour, and Walter Rodney are among the foremost.

Much recreational activity is based upon the festivities that accompany Hindu, Muslim, and Christian holidays. The Guyanese share the passion for cricket that is prevalent throughout the English-speaking Caribbean.

Media

The government has taken nearly complete control over local news media, including the one radio station and the single daily newspaper. Objections against censorship have been on the rise from opposing political and church groups. In 1988 Guyana's first television station was established under government control.

For statistical data on the land and people of Guyana, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

The first human inhabitants of Guyana probably came into the highlands during the first millennium before

Christ. The Warrau Indians may have arrived first, followed by Arawak and Carib tribes. The early communities practiced shifting agriculture supplemented by hunting. Christopher Columbus sighted the Guyana coast in 1498, and Spain subsequently claimed, but largely avoided, the area between the Orinoco and Amazon deltas, a region long known as the Wild Coast. It was the Dutch who finally began European settlement, establishing trading posts upriver in about 1580. By the mid-17th century they had begun importing slaves from West Africa to cultivate sugarcane. In the 18th century the Dutch, joined by other Europeans, were moving their estates downriver toward the fertile soils of the estuaries and coastal mud flats. Laurens Storm van's Gravesande, governor of Essequibo from 1742 to 1772, coordinated these development efforts.

Guyana changed hands with bewildering frequency during the wars (mostly between the British and the French) from 1780 to 1815. During a brief French occupation, Longchamps, later called Georgetown, was established at the mouth of the Demerara; the Dutch renamed it Stabroek and continued to develop it. The British took over in 1796 and remained in possession, except for short intervals, until 1814, when they purchased Demerara, Berbice, and Essequibo, which in 1831 were united as the colony of British Guiana.

Formation of British Guiana

The slave trade was abolished in 1807, when there were about 100,000 slaves in Berbice, Demerara, and Essequibo. After full emancipation in 1838, black freedmen left the plantations to establish their own settlements along the coastal plain. The planters then imported labour from several sources, the most successful group being indentured workers from India. Indentured labourers who had earned their freedom settled in their own coastal villages near the estates, a process that became established in the late 19th century during a serious economic depression caused by competition with European sugar beet production.

Settlement proceeded slowly, but gold was discovered in 1879, and a boom in the 1890s helped the colony. The North Western District was organized in 1889 and was the cause of a dispute in 1895 when the United States supported Venezuela's claims to the territory. Venezuela revived its claims on British Guiana in 1962, an issue that went to the United Nations for mediation in the early 1980s but which remains unresolved.

The British inherited from the Dutch a complicated constitutional structure. Changes in 1891 led to progressively greater power being held by locally elected officials, but reforms in 1928 invested all power in the governor and the Colonial Office. In 1953 a new constitution—with universal adult suffrage, a bicameral elected legislature, and a ministerial system—was introduced.

From 1953 to 1966 the political history of the colony was stormy. The first elected government, formed by the People's Progressive Party led by Cheddi Jagan, seemed so procommunist that the British suspended the constitution in October 1953 and dispatched troops. The constitution was not restored until 1957. The PPP split along racial lines, Jagan leading a predominately East Indian party and Forbes Burnham leading a party of African descendants, the People's National Congress. In the elections of 1957 and 1961, the PPP was returned with working majorities. From 1961 to 1964 severe rioting involving bloodshed between rival blacks and East Indians and a long general strike led to the return of British troops.

To answer the PNC allegation that the existing electoral system unduly favoured the East Indian community, the British government introduced for the elections of December 1964 a new system of proportional representation. Thereafter the PNC and a smaller, more conservative party formed a coalition government, led by Burnham, which took the colony into independence under its new name, Guyana, on May 26, 1966. The PNC gained full power in the general election of 1968, which was characterized by questionable rolls of overseas voters and widespread claims of electoral impropriety. On Feb. 23, 1970, Guyana was proclaimed a cooperative republic within the Commonwealth. A president was elected by the National Assembly, but Burnham retained executive power as prime minister. Burnham declared his government to be socialist and in the later 1970s sought to reorder the government in his favour. In 1978 one of the most bizarre incidents in modern history occurred in Guyana when some 900 members of a cult in a commune called Jonestown committed mass suicide at the behest of their leader, Reverend Jim Jones.

Independence

In 1980, under a new constitution, Burnham became executive president, with still wider powers, after an election in which international observers detected widespread fraud. In the following years Burnham was faced with an economy shattered by the depressed demand for bauxite and sugar and a restive populace suffering from severe commodity shortages and a near breakdown of essential public services. Burnham enforced austerity measures, and he began leaning toward Soviet-bloc countries for support. Burnham died in 1985 and was succeeded by the prime minister, Hugh Desmond Hoyte, who pledged to continue Burnham's policies. In elections held that year Hoyte won the presidency by a wide margin, but once again charges of vote fraud were raised. In the late 1980s the administration, facing worsening financial and economic problems, moved to liberalize the economy.

For later developments in the history of Guyana, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 964, 966, and 974.

BIBLIOGRAPHY

Physical and human geography: General information can be found in WILLIAM B. MITCHELL *et al., Area Handbook for Guyana* (1969), dated but still useful; K.F.S. KING, *Land and People in Guyana* (1968); and FRANCIS CHAMBERS (comp.), *Guyana* (1989), an annotated bibliography. For statistical information, see the *Statistical Digest* (annual). For natural resources, see VINCENT ROTH (comp.), *Handbook of Natural Resources of British Guiana* (1946), and *Notes and Observations on Animal Life in British Guiana, 1907–1941: A Popular Guide to Colonial Mammalia* (1941); and D.B. FANSHAWE, *The Vegetation of British Guiana, a Preliminary Review* (1952). An early botanical study is WALTER E. ROTH (trans. and ed.), *Richard Schomburgk's Travels in British Guiana, 1840–1844,* 2 vol. (1922–23; originally published in German, 1847–48). Studies of the indigenous population include COLIN HENFREY, *A Gentle People: A Journey Among the Indian Tribes of Guiana* (1964; U.S. title, *Through Indian Eyes,* 1965), which also provides a lively travel account; MARY NOEL MENEZES, *British Policy Towards the Amerindians in British Guiana, 1804–1873* (1977); and ANDREW SANDERS, *The Powerless People: An Analysis of the Amerindians of the Corentyne River* (1987). DWARKA NATH, *A History of Indians in Guyana,* 2nd rev. ed. (1970), examines the East Indian population. RAYMOND T. SMITH, *British Guiana* (1962, reprinted 1980), is an outstanding sociological survey, and *The Negro Family in British Guiana* (1962, reissued 1971), is an anthropological classic. Economic aspects of the sugar industry are dealt with in ALAN H. ADAMSON, *Sugar Without Slaves: The Political Economy of Guyana* (1972), on the 19th century; JAY R. MANDLE, *The Plantation Economy: Population and Economic Change in Guyana, 1838–1960* (1973); WALTER RODNEY, *A History of the Guyanese Working People, 1881–1905* (1981); and CLIVE Y. THOMAS, *Plantations, Peasants, and State: A Study of the Mode of Sugar Production in Guyana* (1984). Views of the country's political situation are presented in LEO A. DESPRES, *Cultural Pluralism and Nationalist Politics in British Guiana* (1967); JACQUELINE ANNE BRAVEBOY-WAGNER, *The Venezuela-Guyana Border Dispute: Britain's Colonial Legacy in Latin America* (1984); HENRY B. JEFFREY and COLIN BABER, *Guyana: Politics, Economics, and Society: Beyond the Burnham Era* (1986); and CHAITRAM SINGH, *Guyana: Politics in a Plantation Society* (1988), a survey of postindependence politics.

History: An early history of Guyana is C.A. HARRIS and J.A.J. DE VILLIERS (comps.), *Storm van's Gravesande: The Rise of British Guiana,* trans. from Dutch, 2 vol. (1911, reprinted 1967), extracts from his dispatches written between 1738 and 1772. ALLAN YOUNG, *The Approaches to Local Self-Government in British Guiana* (1958), deals mainly with the 19th century. BRIAN L. MOORE, *Race, Power, and Social Segmentation in Colonial Society: Guyana After Slavery, 1838–1891* (1987), is a history of race relations. THOMAS J. SPINNER, JR., *A Political and Social History of Guyana, 1945–1983* (1984), provides an overview of recent events. CHEDDI JAGAN, *The West on Trial: The Fight for Guyana's Freedom,* rev. ed. (1972, reissued 1980), is a vivid account of preindependence turmoil by a former prime minister. LATIN AMERICAN BUREAU, *Guyana: Fraudulent Revolution* (1984), takes a closer look at Burnham's government.

(B.C.Ri./Ed.)

Gymnosperms

Gymnosperms are a group of vascular plants whose seeds are not enclosed by a ripened ovary (fruit). In 1825 the Scottish botanist Robert Brown distinguished gymnosperms from the other major group of seed plants, the angiosperms, whose seeds are surrounded by an ovary wall. The seeds of many gymnosperms (literally, "naked seed") are borne in cones and are not visible. These cones, however, are not the same as fruits. During pollination, the immature male gametes, or pollen grains, sift among the cone scales and land directly on the ovules (which contain the immature female gametes) rather than on elements of a flower (the stigma and carpel) as in angiosperms. Furthermore, at maturity, the cone expands to reveal the naked seeds. Gymnosperms were considered at one time to be a class of seed plants, called Gymnospermae, but taxonomists now tend to recognize four distinct divisions of extant gymnospermous plants (Coniferophyta, Cycadophyta, Ginkgophyta, Gnetophyta) and to use the term gymnosperms only when referring to the naked-seed habit. Some of the divisions of gymnosperms are not closely related to others, having been distinct groups for hundreds of millions of years. Currently, about 60–70 genera are recognized, with a total of 700–800 species. Gymnosperms are distributed throughout the world, with extensive latitudinal and longitudinal ranges.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 313, and the *Index.*

This article is divided into the following sections:

The gymnosperms: an overview 447
- General features 447
  - Diversity in size and structure
  - Distribution and abundance
  - Importance to humans and ecology
- Natural history 448
- Form and function 449
  - General features
  - Stems
  - Leaves
  - Roots
  - Strobili
- Evolution and paleobotany 450
- Classification 452
  - Distinguishing taxonomic features
  - Annotated classification
  - Critical appraisal

The gymnosperm divisions 452
- Coniferophytes 452
  - General features
  - Natural history
  - Form and function
  - Classification
- Cycadophytes 457
  - General features
  - Natural history
  - Form and function
  - Classification
- Gnetophytes 461
  - General features
  - Form and function
  - Classification
- Ginkgophytes 464
  - General features
  - Form and function
  - Classification

Bibliography 465

THE GYMNOSPERMS: AN OVERVIEW

GENERAL FEATURES

Diversity in size and structure. Among the gymnosperms are plants with stems that may barely project above the ground and others that develop into the largest of trees. Cycads resemble palm trees, with fleshy stems and leathery, featherlike leaves. The tallest cycads reach 19 metres (62 feet). *Zamia pygmaea,* a cycad native to Cuba, has a trunk less than 10 centimetres (four inches) in height. Of the gnetophytes, *Ephedra* (joint fir) is a shrub and some species of *Gnetum* are vines, while the unusual *Welwitschia* has a massive, squat stem that rises a short distance above the ground. The apex is about 60 centimetres in diameter. From the edge of the disk-shaped stem apex arise two leathery, straplike leaves that grow from the base and survive for the life of the plant. Most gymnosperms, however, are trees. Of the conifers, the redwoods (*Sequoia*) exceed 100 metres in height and, while *Sequoiadendron* (giant redwood) is not as tall, the trunk is more massive.

Distribution and abundance. Although since the Cretaceous period (144 to 66.4 million years ago) gymnosperms have been gradually displaced by the more recently evolved angiosperms, they are still successful in many parts of the world and occupy large areas of the Earth's surface. Conifer forests, for example, cover vast regions of northern temperate lands in North America and Eurasia. In fact, they grow in more northerly latitudes than do angiosperms. Vascular plants that occur at the highest altitudes are the gnetophyte *Ephedra.* Land in the Southern Hemisphere is rich in conifer forests, which tend to be more abundant at higher altitudes. Gymnosperms that occupy areas of the world with severe climatic conditions are adapted to conserving water; leaves are covered with a heavy, waxy cuticle, and pores (stomata) are sunken below the leaf surface to decrease the rate of evaporation.

Cycads are distributed throughout the world but are concentrated in equatorial regions. As a natural population, *Ginkgo* originally appeared to have been confined to mountains of southeastern China; extensive artificial propagation has altered this natural distribution. Distribution of gymnosperms in the distant past was much more extensive than at present. In fact, gymnosperms were dominant in the Mesozoic era (245 to 66.4 million years ago), during which time some of the modern families originated (Pinaceae, Araucariaceae, Taxodiaceae).

Importance to humans and ecology. Some of the oldest living things on earth are gymnosperms. Redwoods live for thousands of years, and some specimens of the bristlecone pine, found in the White Mountains of California, approach 5,000 years in age.

Gymnospermous plants are widely used as ornamentals. Conifers are often featured in formal gardens and are used for bonsai. Yews and junipers are often low-growing plants cultivated for ground cover. Conifers are effective windbreaks, especially those that are evergreen. Cycads are used as garden plants in warmer latitudes, and some may even thrive indoors. Their leathery green foliage and sometimes colourful cones are striking. *Ginkgo* is a hardy tree, and although it once approached extinction, it is now cultivated extensively and survives such challenging habitats as the streets of New York City. Some gymnosperms are weedy in that they invade disturbed areas or abandoned agricultural land. Pines and junipers are notorious invaders, making the land unusable.

Horticultural uses

Most of the commercial lumber in the Northern Hemisphere is derived from the trunks of conifers such as pine, Douglas fir, spruce, fir, and hemlock. *Araucaria,* kauri, and *Podocarpus* are important conifers of the Southern Hemisphere used for lumber. The wood is straight-grained, light for its strength, and easily worked. Wood of gymnosperms is often called softwood to differentiate it from the hardwood angiosperms. Wood of angiosperms typically has more kinds of elements than does softwood of gymnosperms. In addition to its use in building construction, gymnospermous wood is used for utility poles and railroad ties. Aromatic wood of cedar is frequently used in the construction of closets or clothes chests and apparently repels cloth-eating moths. Most plywood is gymnospermous. Fibres of conifers make up paper pulp and may occasionally be used for creating artificial silk or other textiles. Conifers are frequently planted in reforestation projects. Conifer bark is often the source of compounds involved in the leather tanning industry. Bark is also used extensively as garden mulch.

Resins

From conifer resins are derived turpentine and rosin. A hardened form of resin from a kauri (*Agathis australis*), called copal, is used in the manufacture of paints and varnishes. Some resins, such as balsam (from hemlock) and dammar (from *Agathis*) are used in the preparation of mounting media for microscope slides. Resins may also have medicinal uses. Many types of amber are derived from fossilized resin of conifers. Commercially useful oils are derived from such conifers as junipers, pines, hemlock, fir, spruces, and aborvitae. These oils serve as air fresheners, disinfectants, and scents in soaps and cosmetics.

Seeds are often food sources. Pine seeds are a delicacy eaten plain or used as a garnish on bakery products. Seeds of *Ginkgo* and cycads may be poisonous unless detoxified. "Berries" (in reality the fleshy cones) of juniper are used to flavour gin.

NATURAL HISTORY

In all living gymnosperm groups, the visible part of the plant body, *i.e.,* the growing stem and branches, represents

From (A,B,E) *The Plant Kingdom* by William H. Brown, © copyright 1935, by William H. Brown, published by Ginn and Company, used by permission of Silver, Burdett & Ginn Inc.; (C,F-J,L-R) T.E. Weier, *Botany: An Introduction to Plant Biology,* copyright © 1982 by John Wiley & Sons, Inc.; (C,H,J,K) redrawn after J.J. Coulter and C.J. Chamberlain, *Morphology of the Gymnosperms,* © 1917 by the University of Chicago; (D) R.E. Torrey, *General Botany for Colleges,* © 1932, renewed 1960, fig. 241A, reprinted by permission of Prentice Hall, Inc., Englewood Cliffs, N.J.

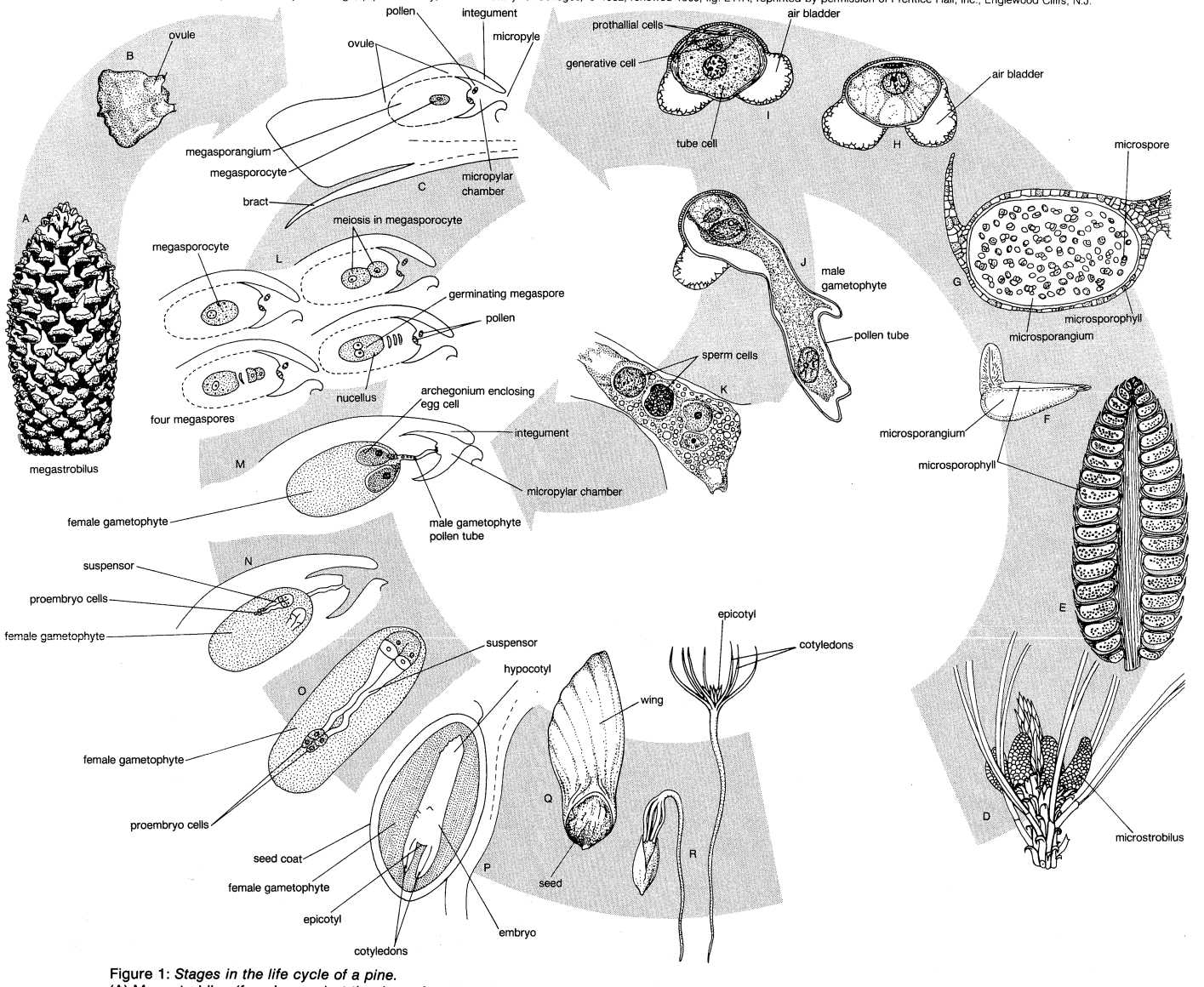


Figure 1: *Stages in the life cycle of a pine.*
(A) Megastrobilus (female cone) at the time of pollination. (B) Ovuliferous scale showing two ovules. (C) Cross section of an ovuliferous scale showing the megasporocyte and the pollen in the micropylar chamber. (D) Microstrobilus. (E) Cross section of a microstrobilus. (F) Scale of a microstrobilus. (G) Cross section of a scale. (H) Winged pollen grain. (I) Development of a male gametophyte in a pollen grain. (J) Pollen tube. (K) Tip of the pollen tube. (L) Stages in megaspore production. (M) Fertilization, showing a female gametophyte with an egg cell and pollen tube. (N) Embryo development. (O) Proembryo. (P) Section of a mature seed showing the seed coats, female gametophyte, and embryo sporophyte. (Q) Winged seed. (R) Germinating seedlings.

Alternation of generations

the sporophyte, or asexual, generation (Figure 1), rather than the gametophyte, or sexual, generation.

In most gymnosperms the pollen cones, called microstrobili, contain reduced leaves called microsporophylls. Microsporangia, or pollen sacs, are borne on the lower (abaxial) surfaces of the microsporophylls. The number of microsporangia may vary from two in many conifers to hundreds in some cycads. Within the microsporangia are cells, called microsporocytes, which undergo meiotic division to produce haploid microspores.

The gametophyte phase begins when the microspore, while still within the microsporangium, begins to germinate to form the male gametophyte. A single microspore nucleus divides by mitosis to produce a few cells. At this stage the male gametophyte (called a pollen grain) is shed and transported by wind or insects.

Ovulate cones, called megastrobili, may be borne on the same plant that bears microstrobili (as in conifers) or on separate plants (as in cycads and *Ginkgo*). A megastrobilus contains many ovuliferous scales, called megasporophylls, that contain megasporangia, within which a single cell (a megasporocyte) undergoes meiotic division to produce four haploid megaspores. Typically three degenerate, leaving one functional megaspore, which is retained within the megasporangium. The female gametophyte begins development within the megaspore. Initial divisions of the megaspore nucleus are mitotic without accompanying divisions of the cytoplasm. As the number of free nuclei multiplies, the megasporangium, integument, and megaspore wall expand. Cell walls eventually develop around the nuclei. At this stage the ovule is ready to be fertilized.

Pollination

Before fertilization can take place, however, the mature male gametophyte (the pollen grain) must be transported to the female gametophyte—the process of pollination. In many gymnosperms, a sticky "pollination droplet" oozes from a tiny hole in the megasporangium (the micropyle) and pollen grains are caught in the droplet. The droplet is then resorbed through the micropyle into the megasporangium. The pollen grain settles on the surface of the megasporangium, where the male gametophyte further develops. A pollen tube emerges from the grain and grows through the megasporangium toward the multicellular egg-containing structure called the archegonium. The egg and sperm continue to mature, the nucleus of the latter undergoing additional divisions resulting in two male gametes, or sperm. (Sperm cells have flagella in cycads and *Ginkgo* but not in any other seed plants.) By the time the pollen tube reaches the archegonium, both the egg and sperm are fully mature, and the egg is ready to be fertilized.

The nuclei of the two sperm are injected into the egg cell; one nucleus dies, and the other unites with the egg nucleus to form a diploid zygote. The nucleus of the fertilized egg begins the development of a new sporophyte generation by dividing a number of times; the resulting multicellular structure becomes the embryo of the seed (Figure 2). Food for the developing embryo is provided by the massive, starch-filled female gametophyte that surrounds it. The time interval between pollination and maturation of the embryo into a new sporophyte generation varies among different groups, ranging from a few months to over one year (in pine, for example). The integument develops into the seed coat, while the female gametophyte is a source of food for the developing embryo during germination.

In some gymnosperms (*e.g.,* cycads, *Ginkgo*) the seed coat (sarcotesta) consists of two layers. In some cycads the sarcotesta is brightly coloured. The sarcotesta of *Ginkgo* seeds is foul-smelling when ripe. Attached to the seed coat in pine and related conifers is a thin membranous wing-like structure, which remains with the seed at its release and serves as a wing that may assist in the distribution of the seed. Members of the order Taxales have a fleshy structure, an aril, surrounding the actual seed. Cones of juniper are fleshy, and the entire fleshy unit drops off or is picked off by birds. Juniper seeds pass through the digestive tracts of birds and are thus distributed effectively.

Maturation of the embryo

At maturity, a gymnosperm embryo has two or more seed leaves (the cotyledons). Cycads, *Ginkgo,* and gnetophytes have two cotyledons in the embryo; pine and other conifers may have several (eight is common; some have as many as 18). Below the attachment point of the cotyledons is the hypocotyl, which emerges through the seed coat during germination, bends downward, and eventually establishes the root system. Above the attachment point of the cotyledons is the epicotyl, the tip of which contains the shoot tip and leaves. In cycads and *Ginkgo* the cotyledons remain within the seed and serve to digest the food in the female gametophyte and absorb it into the developing embryo. Conifer cotyledons typically emerge from the seed and become photosynthetic after digesting and absorbing the food in the female gametophyte.

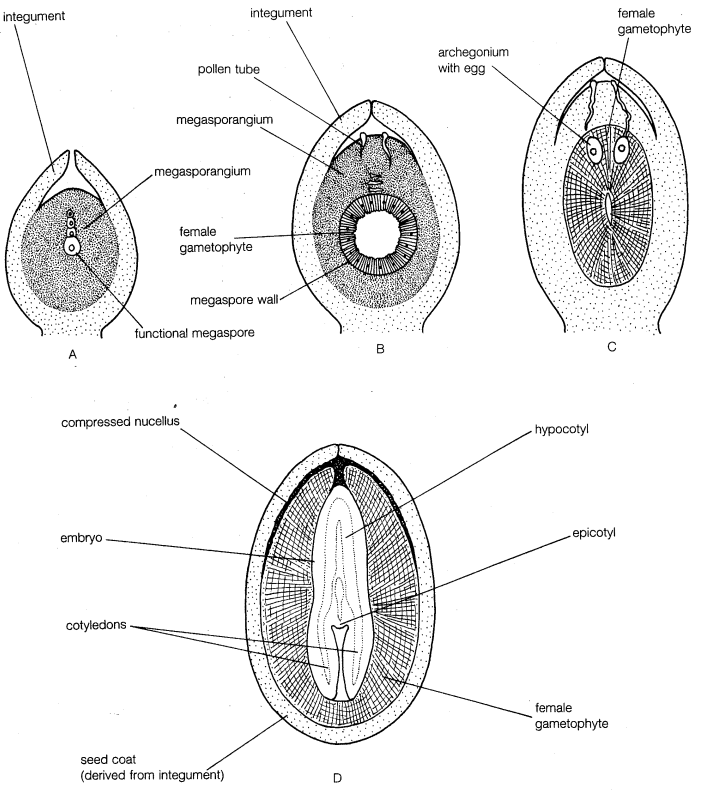


Figure 2: *Stages in the development of the seed in gymnosperms.*
(A) Young ovule at the time of megaspore formation. (B) Development of the megaspore and germination of pollen tubes. (C) Mature ovule at the time of fertilization. (D) Mature seed.

Fig. 30-1 from *Botany* by Peter M. Ray, Taylor A. Steeves, and Sara A. Fultz, copyright © 1983 by Saunders College Publishing, a division of Holt, Rinehart and Winston, Inc.; reprinted by permission of the publisher

FORM AND FUNCTION

General features. The visible part of the gymnospermous plant body represents the sporophyte generation. Typically, a sporophyte has a stem with roots and leaves and bears the reproductive structures. The vascular system contains two conducting tissues, the xylem and phloem. The xylem is a tissue containing nonliving cells whose walls form a conducting system of "pipes" through which water and minerals are conducted from the roots to the shoots. The sturdy nature of the xylem makes it useful in support as well. The phloem, like the xylem, is a conducting tissue; its cells, however, are living and distribute the sugars, amino acids, and organic nutrients manufactured in the leaves to the nonphotosynthetic tissues of the plant. When the plant is actively growing, the phloem may also conduct stored nutrients from the roots to the developing shoots.

Xylem and phloem

Stems. The stems, roots, and branches of vascular plants undergo secondary growth, which takes place from stem and branch growth tissue, called the vascular cambium. Stems of conifers are characteristically woody with a dense mass of secondary xylem. They are usually branched, with basal branches dropping off as the stem elongates, resulting in a main stem that is often tall and straight. The wood is simpler than that of angiosperms; it consists primarily of elongated water- and mineral-conducting cells (tracheids)

in the xylem and living cells that store materials and provide for lateral conduction (vascular rays) in the phloem. The growth tissue of the stem and branches (the vascular cambium) contributes more xylem each growing season, forming concentric growth rings in the wood. Tracheids produced by the vascular cambium early in the growing season are larger, and the walls thinner, than those formed later in the growing season. This results in the characteristic light and dark bands of wood. Some conifers have additional cell types, such as fibres and axially elongated xylem parenchyma cells that store food. Phloem is also simpler than that of angiosperms, consisting of food-conducting cells (sieve cells) and storage cells. Phloem rays traverse the phloem tissue.

Growth rings

Stems of *Ginkgo* are anatomically similar to those of conifers. *Ginkgo* and cedar have two kinds of branches: elongated major branches and dwarf lateral branches. The dwarf shoot bears a cluster of leaves; at the end of the growing season the shoot develops a terminal bud that elongates the following year to produce a new set of leaves. After several years these dwarf shoots develop into short, stubby outgrowths from the stem. Stems of cycads are typically short and squat, although the Australian cycad *Macrozamia hopei* may reach 19 metres. In the centre is a large, fleshy pith surrounded by a cylinder of xylem and phloem. There never is as much secondary vascular tissue as is found in conifers, however. Interspersed among the thin-walled tracheids are abundant vascular rays. The wood, consequently, is not as dense as in conifers.

Leaves. Leaves of gymnospermous plants are extremely variable. Most gymnosperms are evergreen, with leaves lasting more than one growing season. Others are deciduous and drop their leaves at the end of every growing season. Bald cypress (*Taxodium*), larch (*Larix*), and dawn redwood (*Metasequoia*) are examples of deciduous conifers. *Ginkgo* also sheds its leaves in the autumn. Among the conifers, leaves are always simple; that is, the blade is a single unit. Leaves may be small and scalelike (*e.g., Thuja*) or needlelike (*Abies, Picea, Pinus*) or have a broad blade (*Araucaria, Agathis*). In some conifers (*Taxodium*) small branch fragments with numerous needlelike leaves are dropped at the end of the growing season.

Cycad leaves are compound, with thick, leathery leaflets borne in a featherlike (pinnate) arrangement on a main axis (rachis). Produced among the normal photosynthetic leaves of cycads are reduced, pointed, stiff, scalelike leaves called cataphylls. These contribute to the persistent "armour" on the trunk surfaces.

Ginkgo resembles an angiospermous tree in that the woody stem is frequently and irregularly branched and bears broad leaves, which are fan-shaped with dichotomously branched veins. The leaves of *Gnetum* look much like those of dicotyledonous angiosperms. *Ephedra* has small, scalelike leaves.

In certain conifers, such as pine and cedar, leaves are borne on dwarf lateral branches that do not elongate, but are telescoped. In cedar, the dwarf lateral shoots grow forward each year producing a new cluster of needles each season. In pine, however, the number of needles per cluster is small (one to eight) and no more needles are produced on the dwarf shoot after the first year.

Ephedra and *Gnetum* do not produce extensive vascular cylinders; *Ephedra* is shrubby, while some species of *Gnetum* are vines. The stem of *Welwitschia* is somewhat turnip-shaped and does not project very high above the ground. The apex is broad and concave, with leaves and reproductive structures borne along the edges. *Gnetum,* unlike most gymnosperms, has vessels in the xylem. Perforations (pores) at the ends of the conducting elements connect them to adjacent elements.

Roots. Filaments of the fungi called endomycorrhizae live within the cells of the roots of certain gymnosperms, especially conifers. Endomycorrhizal fungi are apparently parasitic, but not destructively so. In cycads, blue-green algae grow in nodules in the roots. These roots may grow opposite to the force of gravity and may form corallike masses on the ground surface, hence the term "coralloid roots." It is thought that these fungi and blue-green algae fix atmospheric nitrogen into a form usable by the plant.

Endomycorrhizae

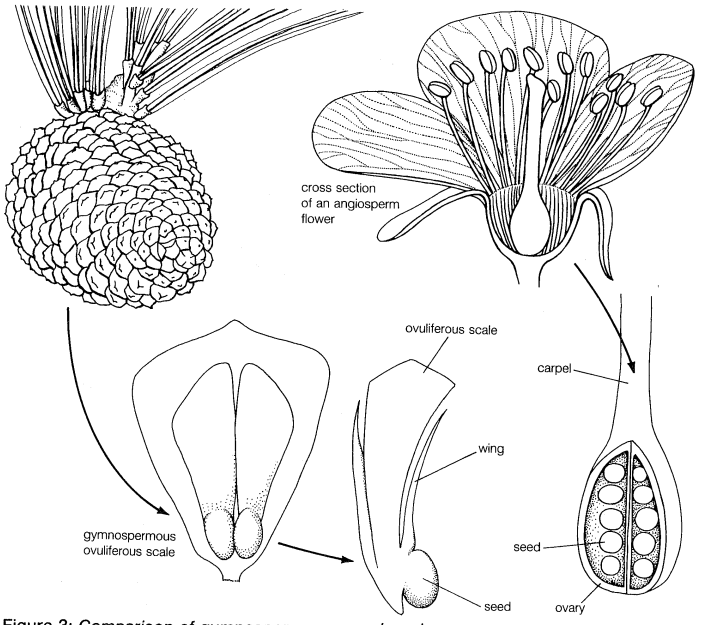


Figure 3: *Comparison of gymnosperm cone and angiosperm flower.*
The seed in gymnosperms is borne exposed (the example above is pine). The angiosperm seed is enclosed within a carpel.

Modified from *Life: An Introduction to Biology* by George Gaylord Simpson and William S. Beck, © 1957, 1965 by Harcourt Brace Jovanovich, Inc., and reproduced with their permission

Strobili. In most conifers the pollen-bearing and ovule-bearing components (the microsporangia and megasporangia, respectively) are borne on the same plant, though separately (monoecious). A pollen-bearing cone, the microstrobilus (Figure 3), consists of a central axis on which are borne, in a close helical arrangement, reduced, fertile leaves (the microsporophylls). On the lower surfaces of the microsporophylls are borne elongated microsporangia; two microsporangia per microsporophyll are common, but some genera have more. The ovulate cone, the megastrobilus, is more complex than the microstrobilus. The megastrobilus bears seeds on flattened dwarf branches, all of the parts of which are fused (ovuliferous scales). Subtending the ovuliferous scale on the cone axis is a reduced scale leaf, or bract. In some conifers the bract is not recognizable because it has been fused to the ovuliferous scale.

In *Ginkgo,* microsporangia and megasporangia are borne on separate trees (*i.e.,* it is dioecious). A *Ginkgo* microstrobilus is borne on a dwarf shoot among the fan-shaped leaves. The microstrobilar axis bears stalked appendages at the ends of each of which are two microsporangia directed downward. A megastrobilus is not recognized as such. Among the leaves of a dwarf shoot on a plant other than one bearing microstrobili are borne elongated, slender stalks, each with a pair of terminal ovules. Usually only one ovule matures into a seed.

Cycads also are dioecious, and all genera bear microstrobili consisting of an axis with microsporophylls inserted in a close, helical arrangement. The microsporophylls are reduced leaves with abaxial sporangia. In the genus *Cycas,* ovules are borne among the edges of the stalk of a reduced leaf with a bladelike region still present. These modified leaves, or megasporophylls, are clustered at the apex of the plant but not arranged in a cone. All other genera of cycads, however, have megastrobili, with the megasporophylls reduced and not leaflike in appearance. Each megasporophyll has a stalk with an expanded distal portion, on the inner face of each of which develop two seeds.

Ephedra, of the Gnetophyta, may have both microstrobili and megastrobili on the same plant, or they may occur on separate plants. The two remaining genera of Gnetophyta, *Gnetum* and *Welwitschia,* are dioecious.

EVOLUTION AND PALEOBOTANY

The first seed plants to have evolved were gymnospermous in the sense that the seeds were naked. The earliest

First seed plants

seedlike bodies are found in rocks of the Late Devonian epoch (374 to 360 million years ago). During the course of the evolution of the seed habit, a number of morphological modifications were necessary. First, all seed plants are heterosporous: two kinds of spores (microspores and megaspores) are produced by the sporophyte. Hence, it is assumed that the ancestors of seed plants must have been heterosporous. Sporangia of plants that do not bear seeds typically lack an integument. The origin of the integument in seed plants was made clear by a study of Early Carboniferous (360 to 320 million years ago) ovules from Scotland. One example, *Genomosperma kidstonii,* consists of an elongated megasporangium with one functional megaspore. Arising from the base of the megasporangium were eight elongated, fingerlike processes that loosely surrounded the megasporangium. In a related species, *G. latens,* these eight fingerlike processes were fused at the base into a cup, with eight free tips. These tips tended to cover the megasporangium rather closely, as opposed to the flared appendages in *G. kidstonii.* Ultimately these fingerlike appendages were almost completely fused into a continuous integument surrounding the megasporangium. A small hole, the micropyle, is left at the apex of the megasporangium where the integument does not quite cover its tip.

In searching for seed-plant ancestors it is necessary to look for a heterosporous type of plant bearing leaves and also having an internal structure similar to that of seed plants. The extinct division Progymnospermophyta provides such an ancestral condition. The best-known progymnosperm is the late Devonian *Archaeopteris,* originally assumed to be a fern with wedge-shaped, subdivided leaflets (pinnules) and sporangia borne on appendages taking the place of pinnules. What was first interpreted as the frond axis was shown to have internal structure like that of *Callixylon,* known as late Devonian stems and wood fragments assumed to be gymnospermous. *Callixylon* wood is like that of many conifers, consisting of tracheids and vascular rays, with closely spaced circular bordered pits on the radial walls of the tracheids. Pits are clustered, separated from other clusters by an area of the wall lacking pits. What were assumed to be pinnae of the frond of *Archaeopteris* are actually branches, and the so-called pinnules are helically arranged leaves. At least some species are known to be heterosporous, hence *Archaeopteris* has many of the features to be anticipated in a seed-plant ancestor.

From progymnosperms such as *Archaeopteris* could have arisen more than one group of gymnosperms. Those with compound leaves (*e.g.,* pteridosperms and cycads) have leaves that would correspond to a flattened branch system of *Archaeopteris.* Those with simple leaves (*e.g.,* conifers) have leaves that are probably the equivalent of the wedge-shaped *Archaeopteris* leaves.

Earliest gymnosperms

The earliest recognized group of gymnospermous seed plants are members of the division Pteridospermophyta (pteridosperms, or seed ferns). These plants originated in the late Devonian and were widespread toward the end of the Paleozoic era (570 to 245 million years ago). Descendants persisted into the Mesozoic. In habit, Paleozoic seed ferns resembled some progymnosperms in that they were small trees with fernlike leaves (the equivalent of a progymnospermous flattened branch) bearing seeds. While Paleozoic seed ferns resembled ferns externally, the internal structure was like that of gymnosperms. Secondary vascular tissues were common in stems of seed ferns. The wood, however, was composed of thin-walled tracheids and abundant vascular rays, suggesting that stems were fleshy like those of cycads. Pteridosperm seeds were very similar to those of cycads. Many were large, with an outer, softer seed coat and a harder, inner seed coat. Within an ovule ready for fertilization was a massive female gametophyte with several archegonia. There has been one report of a pollination droplet in a Carboniferous pteridosperm ovule and a report of a pollen tube emerging from a pollen grain in the micropyle of a seed-fern ovule, suggesting that transport of the sperm through a pollen tube (siphonogamy) was in existence as far back as the Paleozoic. In some pteridosperms the seed was contained within a cupule; some botanists interpret the cupule as a precursor of an angiosperm carpel. Pollen grains, however, landed directly on the micropyle of the ovule. Pollen-bearing organs were variable among the pteridosperms; in many cases the microsporangia were elongated and fingerlike and were produced in clusters or were fused into compound organs. Mesozoic seed ferns are less well defined, and the concept of pteridospermy is used loosely to refer to plants with fernlike foliage bearing seeds.

First cycads

It is generally conceded that from the pteridosperms arose members of the division Cycadophyta. The first cycads appeared in the Permian period (286 to 245 million years ago). Some of these presumed cycads differ from extant members in that megasporophylls were undivided, unlike those of *Cycas,* considered to be primitive among cycads, in which the distal portion of the megasporophyll may be pinnately divided. Other Permian megasporophylls, from China, are more like those of *Cycas.* Cycad remains, especially leaves, are abundant in Mesozoic rocks. For this reason paleobotanists often refer to the Mesozoic era as the "age of cycads." The earliest well-known cycads appear to have had slender stems, sometimes branched, with leaves not borne close together, unlike the situation in extant cycads in which leaves are densely crowded at the apex of the plant. There is evidence that these earliest cycads were deciduous. Megasporophylls of Mesozoic cycads are essentially like those of extant cycads. The megasporophyll of the Triassic *Palaeocycas* is like that of *Cycas.* Jurassic megasporophylls are like those of most other cycads. Extant cycads are now limited in geographic distribution to the warmer parts of the earth.

Coexisting with the cycads during the Mesozoic was another group of gymnosperms, the cycadeoids (division Cycadeoidophyta—sometimes called the Bennettitophyta). Although they are superficially similar in habit to the cycads, with a squat trunk and often pinnately divided leaves, their reproductive structures were different, and their actual relationship is not close. Typically seeds were borne on the surface of a fleshy receptacle. Among the seeds were sterile structures, called interseminal scales, that held the seeds tightly together. Pollen organs were quite similar among the forms in the sense that all had a whorl of modified leaves (microsporophylls) on which were borne compound microsporangia.

Earliest conifers

Conifers (division Coniferophyta) appeared first toward the end of the Late Carboniferous epoch (320 to 286 million years ago). Some of the earliest conifers (class Cordaitopsida) were trees with long, strap-shaped leaves. Trunks were similar to those of extant conifers, with dense, compact wood; small, thick-walled tracheids; and narrow vascular rays. Reproductive axes were slender, bearing narrow bracts in the axils of which were small, budlike shoots with helically arranged scales. On some of the topmost scales were borne elongated microsporangia. Buds on other axes bore ovules instead of microsporangia.

By the late Paleozoic there came into existence another group of extinct conifers, the Voltziales (class Coniferopsida). In general habit they must have resembled some of the extant araucarias (*e.g.,* Norfolk Island pine), with whorled, flattened branches bearing helically arranged, needlelike leaves. Reproductive axes were generally homologous with those of the Cordaitales, but they were more compact, with the bracts on the ovule-bearing axes obscuring the axillary fertile buds. During the end of the Paleozoic and in the early Mesozoic, these axillary buds underwent further transformation. The sterile, non-seed-bearing part became flattened, with the scales fused together. The ovule-bearing portion was situated toward the upper surface (away from the bract). The ovuliferous scale of a conifer seed cone, then, may be interpreted as an axis bearing bracts in the axils of which are modified woody ovuliferous scales derived from lateral buds.

Modern families of conifers began to appear in the Mesozoic era. Members of the Taxodiaceae, the family to which redwoods and bald cypress are assigned, appeared first in the Jurassic period. *Metasequoia,* the dawn redwood, is also a member of this family. Discovered first as fossils in Miocene (23.7 to 5.3 million years ago) deposits, it was assumed to have become extinct until it was discovered growing in Szechwan province in China. Its distribution

in the late Mesozoic and Tertiary (66.4 to 1.6 million years ago) was throughout the Northern Hemisphere. The plant has since been introduced to a variety of places in the world.

During late Triassic times there existed a type of conifer (*Compsostrobus*) that had many features of the Pinaceae. Seed cones had woody ovuliferous scales subtended by bracts with two ovules on the upper surface of each ovuliferous scale. More typical pinaceous remains occurred later in the Mesozoic. Coniferophytes were the dominant vegetation just before the appearance of the angiosperms.

Ginkgo

The division Ginkgophyta, represented today by only one living species, *Ginkgo biloba,* was much more widespread in past ages. Gymnosperms that were presumed to be ginkgophytes existed as far back as the Permian period. In Mesozoic rocks, *Ginkgo* leaves are commonly found throughout the world. The oldest fossil ginkgophytes had leaves much more dissected than the typical *Ginkgo* leaf, resembling more closely the leaves found on new growth in extant ginkgoes.

The fossil record of the division Gnetophyta is obscure, and its origin is not clear. Pollen grains similar to those of *Ephedra* and *Welwitschia* are found as far back as the Permian period. Megafossil remains of possible gnetophytan plants occur in Late Cretaceous (97.5 to 66.4 million years ago) deposits. The plant is unlike any existing one, but venation of the foliage is similar to that of leaves of *Welwitschia.* Pollen grains are typically gnetophytan.

CLASSIFICATION

Distinguishing taxonomic features. Gymnosperms differ from angiosperms most obviously on the basis of the naked-seed habit in the former and the enclosure of seeds within a fruit in the latter. The pollen grain of gymnosperms, when shed from the microsporangium, has more than two cells (three in cycads and four in *Ginkgo* and conifers). Furthermore, gymnosperm pollen lands on the ovule directly, whereas in angiosperms pollen lands on the stigma of a carpel and germinates there, with the pollen tube growing through stigmatic and stylar tissues to reach the ovule. In angiosperm pollen tubes, a total of three cells make up the male gametophyte; gymnosperms have more. The female gametophyte in gymnosperms is much larger than that of angiosperms and serves as the source of food for the developing embryo sporophyte. The female gametophyte of angiosperms consists normally of just a few cells. Both sperm cells in an angiosperm pollen tube are functional, one fertilizing the egg, the other joining with two other nuclei of the female gametophyte. Division of this latter cell forms a multicellular tissue (endosperm) in which food is stored for the embryo. Gymnospermous ovules typically have only one integument; most angiosperms have two.

Annotated classification. Older classifications considered all seed plants to be assignable to a single division, Spermatophyta. The Angiospermae and Gymnospermae were two classes that made up the division. More recent classifications recognize that the characteristic of naked seeds is not important enough to be used to tie all plants with that feature into one group. Classification of gymnosperms now recognizes four separate divisions. Groups marked with a dagger (†) are known only from fossils and have no living members.

†DIVISION PTERIDOSPERMOPHYTA
Late Devonian to Jurassic; seed plants resembling tree ferns with compound, frondlike leaves; seeds and microsporangia borne on the leaves; most stems with secondary vascular tissues.

DIVISION CYCADOPHYTA
Permian to the present; palmlike plants; leaves usually pinnately compound; dioecious; seeds borne in megastrobili with reduced megasporophylls, each bearing inwardly directed seeds (except for the living genus *Cycas*); microstrobili with microsporophylls bearing abaxial microsporangia; 11 extant genera.

†DIVISION CYCADEOIDOPHYTA (BENNETTITOPHYTA)
Triassic (Permian?) to the Cretaceous; cycadlike plants; leaves usually pinnately compound (some entire); ovules borne on the surface of a fleshy receptacle and separated by interseminal scales; microsporangia compound and borne on fingerlike structures fused at the base.

†DIVISION GLOSSOPTERIDOPHYTA
Permian; trees with tongue-shaped leaves with net venation; trunks with compact conifer-like wood; seeds borne on, or associated with, leaves; microsporangia borne on tongue-shaped leaves.

DIVISION CONIFEROPHYTA
Late Carboniferous to the present; woody plants, usually trees, with simple leaves; wood compact; microstrobilus bearing microsporophylls with elongated abaxial microsporangia; seeds borne on megastrobili; ovule with a single integument.

†Class Cordaitopsida
Mississippian to the Permian (or perhaps into the Mesozoic); trees; leaves elongated, strap-shaped; wood compact, coniferlike; fertile shoots slender and elongated with fertile buds borne in the axils of reduced leaves or bracts.

Class Coniferopsida
Late Carboniferous to the present; mostly trees; leaves scalelike, needlelike, or flat and bladelike; wood compact; microstrobili with reduced microsporophylls with abaxial sporangia; megastrobili usually bearing woody ovuliferous scales derived from flattened dwarf branches; seeds borne on the upper surface; 50 genera.

Class Taxopsida
Triassic to the present; trees or shrubs; leaves needlelike; microstrobili with microsporophylls bearing abaxial microsporangia; seeds not in megastrobili but terminating dwarf shoots and surrounded by a fleshy aril; 5 genera.

DIVISION GINKGOPHYTA
Permian to the present; dioecious trees with fan-shaped leaves; bilobed or with more lobes, especially in fossil forms; microstrobili borne among leaves on dwarf shoots; ovules on stalks borne among leaves; 1 extant genus.

DIVISION GNETOPHYTA
An artificial group containing 3 orders: Ephedrales, shrubs with reduced leaves and jointed stems; Welwitschiales, plants with a massive, fleshy stem bearing 2 large, leathery, strap-shaped leaves; and Gnetales, vines, shrubs, or trees with flattened angiospermoid leaves.

Critical appraisal. The classification presented above emphasizes that all gymnospermous plants are not closely related to each other. The characteristic of naked seeds was apparently derived among seed plants more than once. The fossil record indicates that the seed-fern–cycadophyte complex was separate from the conifer line from the very beginning. Plants called pteridosperms may not all belong together; Mesozoic forms were quite different from the Paleozoic ones. In fact, it cannot even be determined in some instances that seeds were actually borne on leaves in the Mesozoic forms. Glossopterids, in a sense, have seed-fern characters in that seeds were borne on fernlike leaves (although they were entire). The wood of glossopterids, however, is more like that of conifers.

Several fossil groups were not included in this classification because their relationships are still obscure. Some of these groups are in the orders Pentoxylales, Vojnovskyales, and Czekanowskiales. (T.De.)

THE GYMNOSPERM DIVISIONS

Coniferophytes

The division Coniferophyta embraces living and fossil plants that usually have needle-shaped, evergreen leaves and seeds attached to the scales of a woody, bracted cone. Three English names—cedar, cypress, and pine—are each applied to unrelated kinds of conifers. Among living gymnosperm divisions, the coniferophytes show little similarity to the Cycadophyta and Gnetophyta but share several vegetative and reproductive traits with the Ginkgophyta. Coniferophytes are most abundant in cool temperate and boreal regions, where they are important timber trees and ornamentals, but they are most diverse in warmer areas, including tropical mountains.

GENERAL FEATURES

Variety among conifers

Diversity of size and structure. The coniferophytes are the most varied gymnosperms (Figure 4). The world's oldest trees are the 4,900-year-old bristlecone pines (*Pinus longaeva*) of desert mountains in California and Nevada. The largest trees are the giant sequoias (*Sequoiadendron giganteum*) of the Sierra Nevada of California, reaching heights of more than 95 metres and weights of at least two million kilograms (4.4 million pounds; compared to 190,000 kilograms for the largest recorded blue whale). Wherever conifers grow, especially in temperate climates, one of these species is usually the tallest tree. In fact, the very tallest trees are the redwoods (*Sequoia sempervirens*) of coastal California, some of which are more than 110 metres tall.

The world's smallest trees probably are also conifers: the natural bonsai cypresses (*Cupressus goveniana*) and shore pines (*Pinus contorta*) of the pygmy forests (adjacent to the towering redwood forests) of the northern California coasts. On the sterile, hardpan soils of these astounding forests, the trees may reach full maturity at under 0.2 metre (7.9 inches) in height, while individuals of the same species on richer, deeper soils can grow to more than 30 metres. Other conifers, such as the pygmy pine (*Lepidothamnus laxifolius*) of New Zealand, the smallest conifer, are always shrubby and may mature as shorter plants (five centimetres in height) than the pygmy cypress, but with greater spread.

(Top left) John Markham—J. Allan Cash Photolibrary, (top right) Grant Heilman—Grant Heilman Photography, Inc., (centre left) © Gilbert Grant from the National Audubon Society Collection—Photo Researchers, (bottom left) G.R. Roberts, Nelson, N.Z., (bottom right) © James E. Eckenwalder

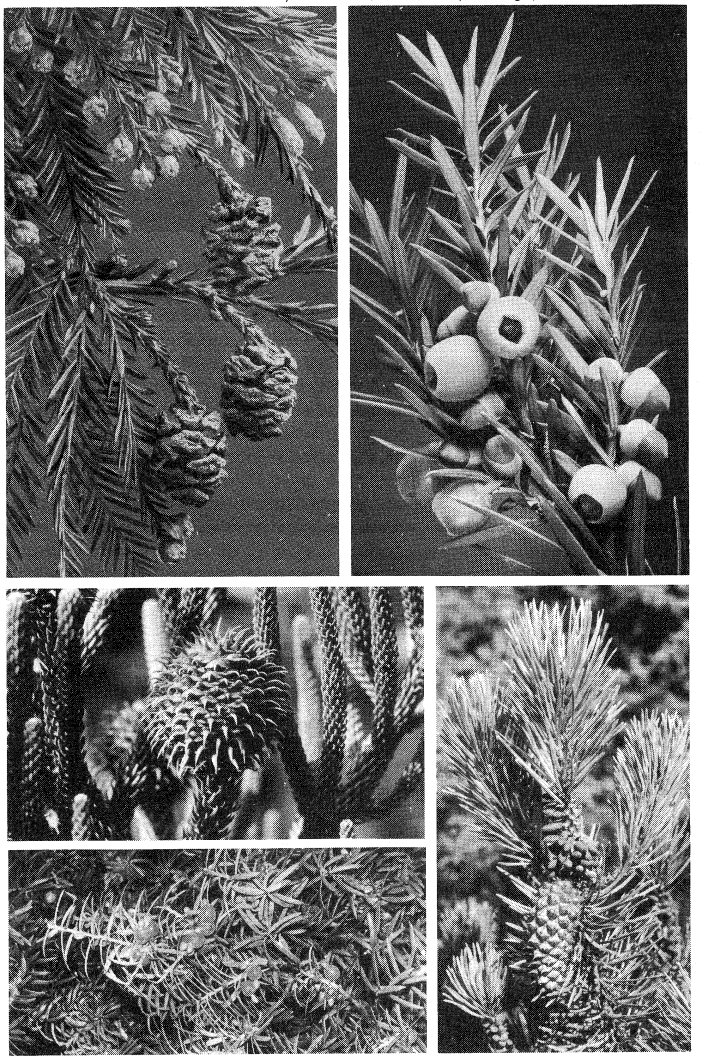


Figure 4: *Representative conifers.*
(Top left) Male (at left and top) and female (right) cones, redwood (*Sequoia sempervirens*); (top right) seeds and fruit, Japanese yew (*Taxus cuspidata*); (centre left) female cone, *Araucaria cunningamii;* (bottom left) fruit, totara (*Podocarpus totara*); (bottom right) male (top) and female (bottom) cones, lodgepole pine (*Pinus contorta*).

Distribution and abundance. Conifers almost cover the globe, from within the Arctic Circle to the limits of tree growth in the Southern Hemisphere. At these extremes, they often form pure stands of one or a few species. The immense boreal forests (or taiga) of northern Eurasia and North America are dominated by just a dozen species of conifers, with even fewer adjunct kinds of hardwoods. The richest north temperate conifer forests are those of mid-latitude mountain systems, where conifers also dominate in numbers. At lower latitudes and moderate elevations are found warm temperate woodlands and forests of pine (*Pinus*), oak (*Quercus,* a hardwood), and juniper (*Juniperus*), which vary in composition and density across Eurasia and North America.

Most tropical conifers are confined to cooler mountain areas where they form solid stands or grow with tropical hardwoods, while a few species inhabit lower elevations. The dammars (*Agathis*), for instance, dominate lowland tropical rain forests in Malaysia, Indonesia, and the Philippines, where they support an important forest industry. Conifers are widespread in southern temperate regions as well, generally with less dominance than in the north. Their greatest diversity is found in the humid portions of the three southern continents, but the largest areas are occupied by semiarid open woodlands of cypress pine (*Callitris*) in Australia and sederboom (*Widdringtonia*) in southern Africa.

Distribution

Conifer species are unevenly distributed. The Eurasian continent is richest in conifers, but every region has its own endemic genera and species. The most widely distributed genera are junipers (*Juniperus*) and pines (*Pinus*), both of which cover the northern continents and extend well into the tropics. Spruces (*Picea*) and firs (*Abies*) are only slightly more restricted. *Podocarpus* is the most widely distributed genus on the southern continents, followed by *Decussocarpus.*

At the other extreme, the most narrowly restricted endemic genera are *Austrotaxus, Neocallitropsis,* and *Parasitaxus* of New Caledonia, an island with the richest conifer flora in the world for its size (14 genera and 44 species). Other highly local genera include *Athrotaxis, Diselma,* and *Microcachrys* in Tasmania, *Fitzroya* and *Saxegothaea* in Chile, *Sequoiadendron* in California, and *Metasequoia* and *Pseudotaxus* in China. Most conifer genera fall between these extremes, with scattered distributions on one or more continents.

Uses in construction

Economic importance. Conifers provide all the world's softwood timber, the major construction wood of temperate regions, and about 45 percent of the world's annual lumber production. Softwoods have always had many general and specialty applications. The original great cedar (*Cedrus libani*) forests of the Middle East were felled to float the warring imperial navies of the ancient world. The same fate later befell the tall North American white pines (*Pinus strobus*) that masted the dominating British navies of the 18th and 19th centuries. Medieval archers drew longbows of the elastic yew wood (*Taxus baccata*). Victims of war and other dead in East Asia have been buried from earliest recorded times in coffins of sugi (*Cryptomeria japonica*) and sanmu (*Cunninghamia lanceolata*), relatives of the equally decay- and termite-resistant redwood (*Sequoia sempervirens*) and bald cypress (*Taxodium distichum*). In the family Cupressaceae are the fragrant cedars. Some are still used to line the chests that protect fine fabrics and furs against insects, but the wonderful fragrance of sharpened lead pencils has disappeared as eastern red cedar or pencil cedar (*Juniperus virginiana*) has been superseded by tropical hardwoods.

The domination of softwoods in lumber construction in northern temperate regions has been further extended by composite products such as plywood, particleboard, and chipboard. Other processed softwood products include paper and plastics derived from chemically treated wood pulp of spruces (*Picea*), tannins from the bark of hemlock (*Tsuga canadensis*), and naval stores (including turpentine) from many pines. Foods and beverages from conifers

include pine nuts and gin, which is flavoured with juniper berries. Canadian balsam, from resin blisters on the bark of the balsam fir (*Abies balsamea*) of northeastern North America, is used as a mounting medium for microscopic preparations.

Conifers are popular ornamentals in parks, cemeteries, and other public places, as well as around private homes and gardens. Although few species are grown indoors as houseplants, the traditional Christmas tree of western Europe and North America brings the fragrance and freshness of the forest into homes during the depth of the northern winter.

NATURAL HISTORY

All conifers share a typical seed-plant life cycle with a long-lived, dominant, photosynthetic, diploid sporophyte and a reduced, transient, dependent, haploid gametophyte (Figure 5). All phases of this general life cycle vary among conifers.

Sporophyte phase. The sporophytes of all conifers are trees or shrubs. They have a life span that ranges from a few decades to more than 4,000 years. The ecological role and way of life of this sole photosynthetic phase of the conifer life cycle varies with the size, form, and habitat of each species. Where conifers are ecologically dominant, as in the boreal and montane forests of the Northern Hemisphere, including the Douglas fir forests of western North America, they may make up 90 percent or more of all the living matter and they contribute greatly to the biosphere through photosynthesis.

Importance of fires

Fires play an important role in many conifer forests. Most conifers contain highly flammable resins. The flammability of these trees increases during hot, dry fire weather, when the water content of the living needles is drastically reduced. Few adult conifers can withstand a conflagration. The giant sequoia (*Sequoiadendron giganteum*) is an outstanding exception because it has insulating bark more than 50 centimetres thick.

Despite their susceptibility to fires, many conifers actually depend on these apparent catastrophes for regeneration.

From G.S. Allen and J.N. Owens, *The Life History of the Douglas-Fir* (1972), Canadian Forestry Service, © Minister of Supply and Services Canada

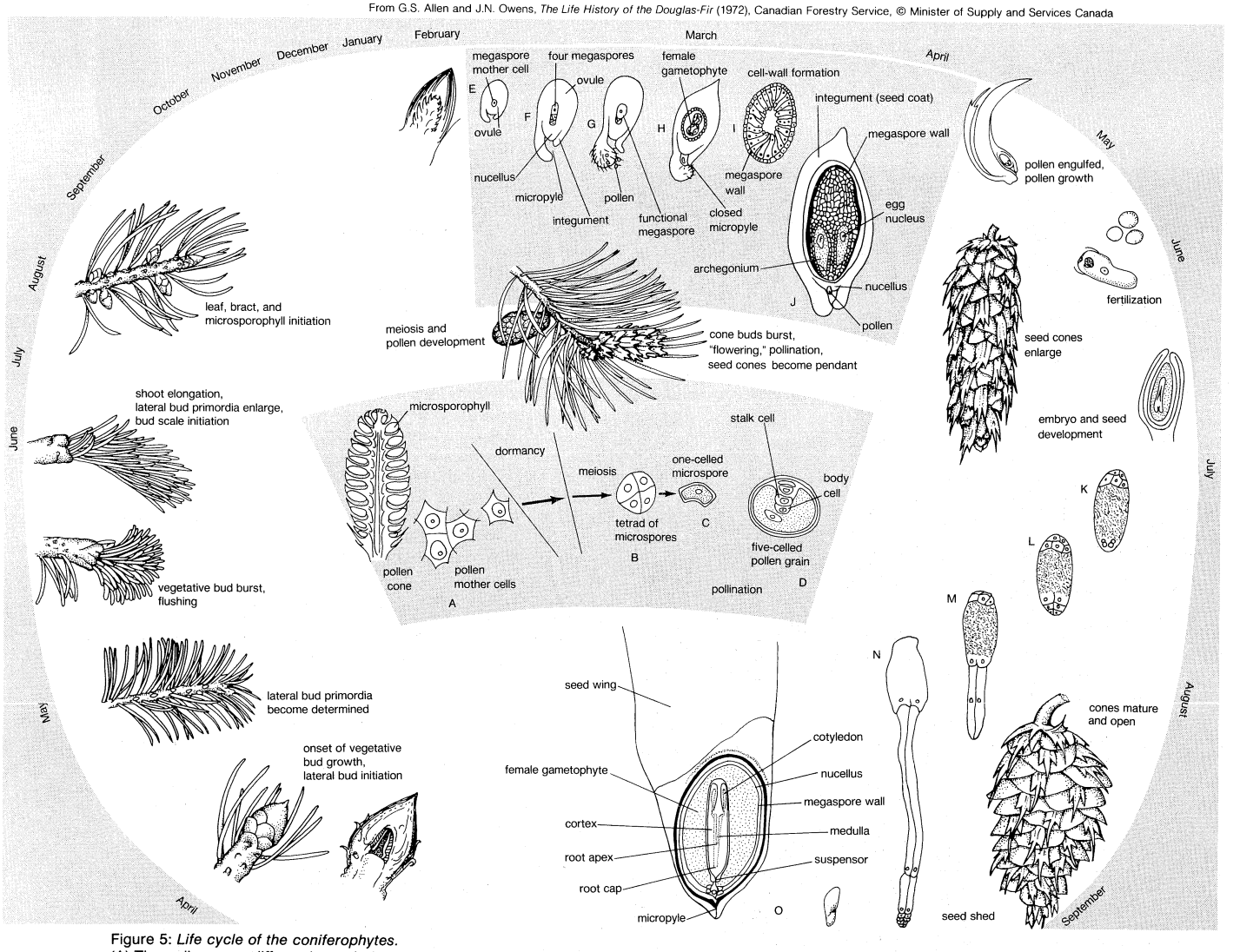


Figure 5: *Life cycle of the coniferophytes.*
(A) The pollen cone differentiates during summer and fall. Pollen mother cells begin meiosis in October; meiosis becomes arrested during winter and resumes in February. (B) Pollen mother cells form tetrads of single-celled microspores. (C) One-celled stage; microspore undergoes successive stages, culminating in (D) mature, five-celled pollen grains, which usually are shed early in April (pollination), after which pollen divides into two male cells (gametes). The stages of ovule and female gametophyte development occur during March, April, and May. (E) An ovule. (F) Meiosis occurs and four haploid megaspores are formed. (G) Three megaspores degenerate and one enlarges; pollination occurs. (H) The stigmatic tip grows inward, drawing pollen in; the megaspore divides without cell-wall formation. (I) Cell-wall formation after hundreds of nuclei have formed. (J) The large egg cell fills the archegonium before fertilization. (K) After fertilization, nuclei divide. (L) Cell walls form. (M) Primary suspensor cells elongate, forcing embryo (lower) tier through archegonial wall. (N) Elongation of suspensor cells and embryonal tube cells pushes embryo deep into female gametophyte tissue. (O) A section through the seed shows a dormant embryo embedded in the female gametophyte.

In such fire-dependent forests (including giant Sequoia groves, Douglas fir forests, boreal forests, low latitude pine forests, and Australian cypress pine woodlands), the dominant conifers are unable to regenerate among the more shade-tolerant species that grow up around them with time. Fires clear the understory to bare soil, stimulating the germination and establishment of seeds. Most of these species have cones that protect the seeds from the worst of the fire and then open to scatter them on the ash-fertilized seed bed. In the boreal forests of Eurasia and North America or in the pine forests of the southeastern United States, the cycle of fires and regeneration may repeat itself every 80 to 120 years, while the giant Sequoia is a fire-dependent successional species with individuals that can live for more than 3,000 years.

Contribution of flooding

At the other extreme are flooded swamp forests of bald cypress (*Taxodium*) in the southeastern United States and *shuaisuong* (*Glyptostrobus*) in southeastern China. Reproduction of these trees is as attuned to flooding as that of fire species is to scorched earth. Their seeds have air and resin pockets that allow them to float away to slightly raised areas revealed by receding floodwaters.

Without extremes of fire and flood, mesophytic species living in temperate and tropical mixed forests may dominate or grow scattered among other trees. The conifers with broad, flat blades rather than needle leaves almost all live in moist forests, as do most species whose seeds have fleshy structures that attract birds or small mammals.

Gametophyte phase. The gametophytes of conifers, like those of other seed plants, live out their brief, nonphotosynthetic lives almost entirely within the spore wall. All of their nutrition is derived from the parent sporophyte. The female gametophyte is never released from the tree until the seed matures. The male gametophyte is briefly separated from the sporophyte when pollen is released into the wind. These pollen grains contain an immature male gametophyte enclosed and dispersed in the microspore wall (Figure 5). In the Pinaceae, three successive divisions of the microspore produce a four-celled pollen grain within the microsporangium. It has two tiny prothallial cells (the last body remnants of the old free-living gametophyte), a tube cell, and a generative cell. After pollination, the tube cell develops the pollen tube and the generative cell divides to form a sterile cell and a spermatogenous cell. Prior to fertilization, the spermatogenous cell divides again to produce two male gametes. Other conifers share the later phases of male gametophyte development with the Pinaceae, but vary in the number of prothallial cells, from none in *Cephalotaxus, Sciadopitys,* Cupressaceae, and Taxaceae to as many as 40 in *Agathis* of the Araucariaceae, which has the most complex male gametophytes among the seed plants. Unlike the ovule (megasporangium), which houses a solitary female gametophyte, each microsporangium produces hundreds or thousands of pollen grains.

The pollen grain

The female gametophytes of conifers are more massive and complex than their male counterparts and basically resemble gametophytes of *Ginkgo* and the cycads. The life history of the female gametophyte begins with a protracted series of free nuclear divisions in the megaspore. At the end of these divisions, there may be up to 2,000 nuclei in a thin layer of cytoplasm pressed against the megaspore wall by a giant central vacuole. Cell walls then form between adjacent nuclei and gradually extend into the central vacuole until the entire gametophyte is filled with radially elongated alveolar cells that are equivalent to the prothallial cells of the pollen grain. This stage is followed by the appearance of archegonia at the micropylar end of the ovule. One to eight archegonia are usual in the female gametophyte of conifers, but there may be up to 200 in some species, each of which can produce an embryo if fertilized. Each archegonium has a single huge egg cell capped by a ventral canal cell and separated from the micropylar surface of the gametophyte by a short neck made up of one or two layers of neck cells. The archegonial end of the female gametophyte usually protrudes from the megaspore wall, which might otherwise prevent pollen tube penetration and fertilization.

Megaspore maturation

Pollination. Like *Ginkgo,* but unlike at least some cycads and gnetophytes, all conifers are pollinated by wind. Pollen may be produced in enormous quantities, particularly by species of true pine (*Pinus*), which can blanket the surface of nearby lakes and ponds with a yellow scum of pollen. The pollen grains of many Pinaceae and Podocarpaceae have air bladders, which orient them in a pollination droplet exuded by the ovules so that, when this droplet is withdrawn back into the ovule, the pollen tube will penetrate the nucellus to the archegonium. The pollen grains of families that lack prothallial cells are more or less spherical, lack air sacs, and can extend a pollen tube anywhere on their surface so that precise orientation is unnecessary. Some conifers lack a pollination droplet mechanism. Douglas fir pollen grains land on an enlarged, stigmalike growth of the micropyle, whence the pollen tubes grow into the nucellus and archegonium. The pollen grains of the Araucariaceae land on the scales of the female cone, and the pollen tubes reach the micropyle by burrowing into the cone scales.

Fertilization and embryogeny. After passing through the nucellus, the pollen tube presses between the neck cells of the archegonium and ruptures to release the tube nucleus, sterile cell, and the two male gametes (sperms). The ventral canal cell seems to help the male gametes enter the egg. One of the sperms fertilizes the egg nucleus to form the zygote, the first cell of the new sporophyte generation.

The zygote

The conifer zygote has fewer free nuclear divisions than do *Ginkgo* or the cycads. While many divide twice to form four free nuclei in the centre of the egg cytoplasm, there may be from zero to six free nuclear divisions. The nuclei usually move away from the micropyle, and cell-wall formation accompanies further cell divisions. The embryo develops and is fed by the nutritive tissue of the female gametophyte. The embryo rapidly enlarges at the expense of the maternal tissue and initiates typical sporophytic organization, consisting at maturity of a single axis with a root apex at one end and a shoot apex at the other, surrounded by two to eight cotyledons.

Germination. The mature seed consists of the dormant embryo embedded in remnants of the female gametophyte and megasporangium (nucellus) and surrounded by a seed coat. The seed coat of conifers is similar to that of other gymnosperms, developing from an integument with three distinct layers. Only the hard middle stony layer is evident in most conifers, protecting the embryo between seed release and germination. The outer fleshy layer is most prominent in those conifers, such as Cephalotaxaceae and some Taxaceae, whose seeds are dispersed by animals. The inner fleshy layer functions in the early development of the ovule, but persists only as a thin, papery membrane in the mature seed.

Germination proceeds immediately upon dispersal to a suitable site in many tropical conifers, but most cool temperate species require a winter period of cold, moist stratification before they will germinate. After the embryo absorbs water, a seedling root breaks through the seed coat and turns down into the soil.

The stem below the cotyledons elongates and lifts them above the ground. As the cotyledons begin to photosynthesize, they produce the energy needed for the early growth of shoots. The seedling shoot is densely clothed with needle-shaped juvenile leaves for a varying time until adult foliage forms; some cedars of the family Cupressaceae produce their first flattened side shoots within just a few nodes, while the longleaf pine (*Pinus palustris*) of the southeastern United States remains in a juvenile "grass" stage for years.

FORM AND FUNCTION

The basic organization of the conifer sporophyte resembles that of other seed plants. The four main organs (stems, leaves, roots, and sporangia) are all usually distinct from one another and have well-defined physiological functions. The considerable variation that occurs in these organs are commensurate with the varied environments in which the different species grow.

Stem. Stems raise the photosynthetic leaves into the light and provide a channel for nutrients between the leaves and the roots. Most of the diameter of mature conifer

Stem function

stems consists of secondary xylem (wood) produced by the vascular cambium, a permanent cylinder of dividing cells that lies just inside the bark. Because the growth of most conifers is cyclic, the wood generally consists of distinct growth rings. These are delimited by the juxtaposition of small dark cells from the end of a growing season with larger, lighter cells that mark resumption of growth. In temperate conifers, these rings correspond to annual growth flushes since no wood is formed during winter dormancy. Tropical species may lack this correspondence unless their habitat has strong seasonal variation in rainfall. Otherwise, they may have more than one growth ring in a year, each accompanying a new flush of leaves and branches.

The wood of conifers is generally more uniform and simpler in structure than that of flowering plants. One type of cell, the tracheid, serves both to transport water and to support the trunk so that conifers lack the more textured wood associated with the mixture of vessel elements and fibres in hardwoods. The wood may have longitudinal resin canals lined with living cells, but most of its living cells are found in the rays that extend horizontally from the centre of the stem to the vascular cambium. The pits, the tiny thinnings that connect adjacent wood cells, are quite varied among conifer families and genera and are one of the chief features used to identify conifer woods.

Bark

The bark that clothes the trunks may be thin, smooth, and flaky, peeling annually to reveal fresh bright colours, as in the lacebark pine (*Pinus bungeana*) of China and the tecate cypress (*Cupressus guadalupensis*) of southern and Baja California, or it may accumulate in broad, colourful plates, as in the ponderosa pine (*Pinus ponderosa*) of western North America, or as thick, fireproof, fibrous ridges on the giant sequoias (*Sequoiadendron giganteum*). The practiced eye can distinguish different species of co-occurring conifers by their bark alone.

Shoots

Some conifers have transient special determinate twigs called short shoots that carry most of the photosynthetic leaves. In bald cypress (*Taxodium*) and dawn redwood (*Metasequoia*), these short shoots look like double-sided combs and are shed each fall, to be followed by new ones in the same spots on the branches when growth resumes the following spring. The short shoots of pine (*Pinus*), the bundles of two to five (rarely one or eight) needles, are retained for up to 20 years, but are finally shed and almost never grow out as branches. In contrast, the peglike short shoots of larch (*Larix*) and true cedar (*Cedrus*), like those of *Ginkgo,* are permanent, elongating slowly and producing new needles each year. The flattened sprays of branchlets of cedars in the family Cupressaceae may act like short shoots, falling intact after a relatively brief life, or they may provide the framework for further branching.

The strangest conifer shoots are the phylloclades that give the celery pine (*Phyllocladus*) its name. These flattened branches look like fern fronds, are green, and are the main photosynthetic organs of the tree. The true leaves are tiny scales that contribute little to overall food production.

Leaves. Leaves are specialized photosynthetic organs. The varied leaves of conifers are attached singly along the stems in a helical pattern or in opposite pairs or trios. Many Cupressaceae and a few other conifers have reptilian scale leaves only a few millimetres long. Diverse needle- and claw-shaped leaves range in length from about one centimetre in many conifers to more than 30 centimetres in some species of pine. Broad, flat, oblong blades up to 30 centimetres long occur in *Agathis* and some species of *Decussocarpus* from East Asia, and the monkey puzzle tree (*Araucaria araucana*) of Chile has hefty triangular wedges. The notched needles of the Japanese umbrella pine (*Sciadopitys verticillata*) are the oddest leaves among living conifers. They could be needlelike phylloclades or a pair of longitudinally fused needles. The largest coniferophyte leaves were those of the extinct genus *Cordaites,* with great paddle- or strap-shaped leaves up to one metre long and 15 centimetres wide.

Leaf function

Most conifer leaves, whatever their shape, minimize water loss. The reduced surface area of the scale- to needle-shaped leaves is an obvious example, but even the broader forms often have a thick, waxy coating that makes them waterproof. The gas-exchange openings of the leaves (stomates) are usually confined to a pair of narrow bands on the undersurface and are deeply sunken into chambers that separate them from direct contact with the dry air surrounding the leaf.

Roots. Roots gather water and mineral nutrients from the soil and anchor and support the above-ground portions. Most conifers have rather shallow, if wide-spreading, root systems, making the trunks highly susceptible to wind and surface disturbance. Even the largest conifers are no exceptions, and many of the individual giant sequoias (*Sequoiadendron giganteum*) in national parks in California are ringed by fences to reduce damage to the roots by the footsteps of millions of admiring visitors. The roots are the least-studied parts of the conifers but appear to be relatively uniform throughout the group. The specialized roots (called haustoria) by which the only parasitic conifer, *Parasitaxus ustus,* attaches to the roots of its conifer hosts are an exception, but the oddest root structures are the "knees" of bald cypresses (*Taxodium distichum*), conical masses of woody tissue that emerge from the swamp waters around each tree. They probably help aerate the roots, which need oxygen to survive and function.

Associations

The fine feeding roots of conifers, like those of many flowering plants, do not work alone. They get a boost in their work by associating with specialized fungi whose structural filaments (hyphae) intermingle with them to form mycorrhizae. There are two distinct types of mycorrhizal associations among the conifers. The majority of species have vesicular-arbuscular mycorrhizae, called endomycorrhizae because the fungal hyphae actually penetrate the cells of the roots. All of the Pinaceae, and only the Pinaceae, have the other kind of root symbiosis, called ectomycorrhizal because the fungi sheath the rootlets and hyphae pass between the outer root cells without penetrating them. Each year, new roots grow out from the sheath and are only recolonized when the fungi later resume active growth. Ectomycorrhizal fungi reproduce through the attached mushrooms that are seen sprouting in pine forests, whereas endomycorrhizal fungi do so underground.

Strobili. The sporangia of vascular plants are technically asexual, but in the seed plants, because the gametophytes are wholly dependent upon the sporophyte and the female gametophyte even remains within the megasporangium, sexual terminology is extended to the sporophyte and sporangium-bearing organs. In all coniferophytes, the organs containing microsporangia ("male") are separate from those bearing megasporangia ("female"), and in *Cephalotaxus,* some junipers (*Juniperus*), and the family Taxaceae, these are found on different individuals.

The microsporangia of all conifers are attached to the scales of a simple pollen cone, or microstrobilus. The pollen cones usually consist of thin, parchmentlike scales (microsporophylls), each carrying two or more microsporangia on the lower surface. The number of scales and their size is quite variable, so that the overall length of the microstrobilus ranges from about two millimetres in some cypress (*Cupressus*) species to more than 20 centimetres in some *Araucaria* species.

Variations among megasporangia

Wide variations in the female (megasporangiate) reproductive structures among the conifers are the main basis for their classification. Most living conifers have a seed cone that is interpreted as a compound strobilus; each cone scale, inserted in the axil of a bract, is equivalent to an entire simple pollen cone. Fossil evidence shows how each ovule-bearing dwarf shoot of ancestral conifers was reduced and fused to form a single cone scale. Like the leaves, the bracts and scales are spirally arranged or occur in pairs or trios on the axis, and modern conifers have at least some fusion between each bract and its scale. The bracts and scales, or combined scales, vary in texture from woody to leathery, or even fleshy in bird-dispersed junipers (*Juniperus*) and family Podocarpaceae. The number and size of cone scales varies widely in the conifers, leading to seed cones that range from three millimetres long and less than a gram in *Microbiota* of the Amur region of the Soviet Union to more than 40 centimetres long in the sugar pine (*Pinus lambertiana*) of California

and more than 2.2 kilograms in some araucarians and the coulter pine (*Pinus coulteri*) of California.

The megasporangiate strobili of *Cephalotaxus* and most Podocarpaceae have the same basic structure as other conifer cones, but are so reduced and dominated by their much larger seeds that they do not look like cones. Even greater modification in the family Taxaceae has completely eliminated any trace of strobilar organization, and the solitary seeds sit at the tip of a short branch in a fleshly aril, a cup-shaped outgrowth of the seed stalk.

CLASSIFICATION

Distinguishing taxonomic features. Extant conifers differ from other gymnosperms in combining simple pollen cones with compound seed cones (or solitary terminal seeds in family Taxaceae). Although not possessed by all species, only conifers have needle leaves (of a variety of shapes) and pollen with bladders. Some other features, although not exclusive to coniferophytes, are also more common in them than in other gymnosperms. These include flattened, winged seeds (also in *Welwitschia*), scalelike foliage leaves (also in *Ephedra*), and the growth habit of a normal tree or shrub (also in *Ginkgo*).

Annotated classification. With more than 50 genera and 550 species, classification of the extant coniferophytes remains controversial. Disagreements exist throughout the classification, so that the numbers of orders, families, genera, and species are all disputed. The classification outlined here reflects current opinion for living conifers but simplifies extinct groups because the number of families to be recognized among the fossils is so uncertain. Extinct groups are indicated by a dagger (†).

†CLASS CORDAITOPSIDA

Paleozoic; strap-shaped leaves, up to 1 metre long, much larger than those of true conifers; both pollen and seed cones were compound and open, each bract with an axillary branch bearing numerous scale leaves surrounding pollen sacs or ovules; generally considered the ancestors of the Coniferales.

CLASS CONIFEROPSIDA

Contains coniferophytes; ovules attached to the scales of a condensed compound seed cone; families defined by seed-cone structure.

†Families Walchiaceae and Voltziaceae

Paleozoic and Mesozoic; show many stages in the transformation of the seed-bearing dwarf shoots of cordaiteans into the unified, flattened seed scales of modern conifers; foliage resembled that of araucarians; include *Walchia, Voltzia,* and *Voltziopsis.*

†Family Cheirolepidiaceae

Mesozoic; scales shed from the cone together with the seeds; large bracts remain attached to the axis in a semblance of a complete cone; distinctive pollen, called *Classopollis;* foliage resembles that found in the modern Cupressaceae; great variety of life-styles.

Family Pinaceae

Largest and most widespread and abundant modern conifer family in the Northern Hemisphere; woody, usually thin, cone scales carry two winged seeds and are fused to the bracts only at their bases; bracts usually hidden by the scales in mature seed cones but may be prominently exserted in some species; leaves are most often needlelike and spirally arranged, either singly or in clusters; pine (*Pinus*), spruces (*Picea*), firs (*Abies*), and larches (*Larix*) are all found across the Northern Hemisphere, while Douglas firs (*Pseudotsuga*) and hemlocks (*Tsuga*) are restricted to North America and Asia, *Cathaya, Keteleeria,* and *Pseudolarix* are restricted to China, and the true cedars (*Cedrus*) occur from Morocco to the Himalayas; 10 to 12 extant genera; about 200 species.

Family Araucariaceae

From Triassic; massive seed cones with a single large seed on each cone scale; highly reduced scales completely fused to the much larger bracts; species of *Araucaria* have branches densely clothed with prickly, spirally arranged, clawlike-to-wedge-shaped leaves; dammars (*Agathis*) have well-separated, oppositely arranged oval or oblong leaves; found in South America, Southeast Asia, and Australasia; 2 extant genera and 30 to 40 species.

Family Sciadopityaceae

Umbrella pine (*Sciadopitys verticillata*) usually included in the Cupressaceae, but recent work confirms its isolation from that family; seed cones superficially resemble those of the giant sequoia (*Sequoiadendron giganteum*), but the equal-sized scales and bracts fused for only about two thirds of their length, each having 5 to 9 seeds; foliage consists of whorls of about 15 to 20 double needles separated by stem segments with spirally arranged, nonphotosynthetic scale leaves; endemic to Japan.

Family Cupressaceae

Although species of this family are traditionally divided between two families, Cupressaceae for the cypresses (*Cupressus*) and similar genera and Taxodiaceae for the much more varied genera allied to the bald cypress (*Taxodium*) and redwood (*Sequoia*), present evidence shows that all belong to a single family; scales of seed cone intimately fused to the bracts; scale complexes vary in texture, shape, and arrangement on the cone; most have 3 to 5 seeds per scale, but the number ranges from 1 to about 20; leaves vary in shape from scales to clawlike or needlelike and are spirally arranged or in opposite pairs or whorls of 3; several genera, usually referred to as cedars (such as *Calocedrus, Chamaecyparis, Libocedrus,* and *Thuja*), have flattened sprays of frondlike branches closely covered with scale leaves; considerable diversity in both the Northern (18 genera) and Southern (11 genera) hemispheres; 50 or more species of junipers (*Juniperus*) are widespread, exceeding even the pines in their coverage of the Northern Hemisphere; other genera include *Athrotaxis, Callitris, Cryptomeria, Cunninghamia, Diselma, Fitzroya, Metasequoia, Microbiota, Neocallitropsis, Sequoiadendron,* and *Widdringtonia.*

Family Podocarpaceae

Seed cone is reduced, with 1 to few highly modified, brightly coloured, fleshy scales, each called an epimatium and surrounding a single seed; most with spirally arranged, yewlike needles, but scale leaves and opposite, broad, oblong blades (up to 30 cm long and 5 cm wide) are also found; about 130 species in 6 to 18 genera, including *Decussocarpus, Lepidothamnus, Microcachrys, Parasitaxus, Phyllocladus, Podocarpus,* and *Saxegothaea.*

Family Cephalotaxaceae

Seed cones highly modified with a few opposite pairs of small bracts, each with a greatly reduced scale remnant strongly dominated by a pair of ovules; only 1 ovule develops into a large seed with a fleshy seed coat; leaves are large yewlike needles carried in opposite pairs; found in East Asia, the plum yews (*Cephalotaxus*) are the second smallest and most local extant conifer family; 1 genus and 4 to 7 species.

Family Taxaceae

Solitary ovules borne at the end of a dwarf shoot bearing densely spiraled scale leaves; mature seeds surrounded by a fleshy aril and may also have fleshy seed coats; *Taxus* seed is highly toxic; pollen cones and flattened, needlelike leaves are more like those of other conifers than are their seeds; widely distributed in the Northern Hemisphere, while *Torreya* is found in restricted areas of both North America and East Asia, *Amentotaxus* and *Pseudotaxus* are localized in China, and *Austrotaxus* is endemic to New Caledonia; 5 genera and 16 species often segregated into an order separate from the Coniferales because of absence of seed cone. (J.E.Ec.)

Cycadophytes

Although some botanists prefer to restrict the term cycadophyte to the members of the division Cycadophyta, three groups of primitive seed plants are discussed here, of which the seed ferns (division Pteridospermophyta) and cycadeoids (division Cycadeoidophyta) are represented only by extinct forms. A third order, Cycadales (cycads), is today represented by 10 or 11 living genera and some 130–160 species.

The cycadophytes encompass a diverse collection of mostly extinct primitive seed plants that probably had their origins among the progymnosperms of the Devonian period (408 to 360 million years ago), possibly among a primitive, long-extinct group of non-seed-bearing plants, the Aneurophytaceae, in which disposition of fertile structures and patterns of branching bear some resemblance to those of seed ferns.

GENERAL FEATURES

Diversity. *Seed ferns.* A number of lines of seed-bearing gymnospermous plants are discernible among fossil plants of the late Paleozoic era (570 to 245 million years ago) and early to middle Mesozoic era (245 to 66.4 million years ago). Among them a rather loose assemblage of forms, referred to as seed ferns, or pteridosperms, is well represented. The Carboniferous period (360 to 286 million years ago) especially has been called the "age of ferns" because of the abundance of fossilized fernlike leaves.

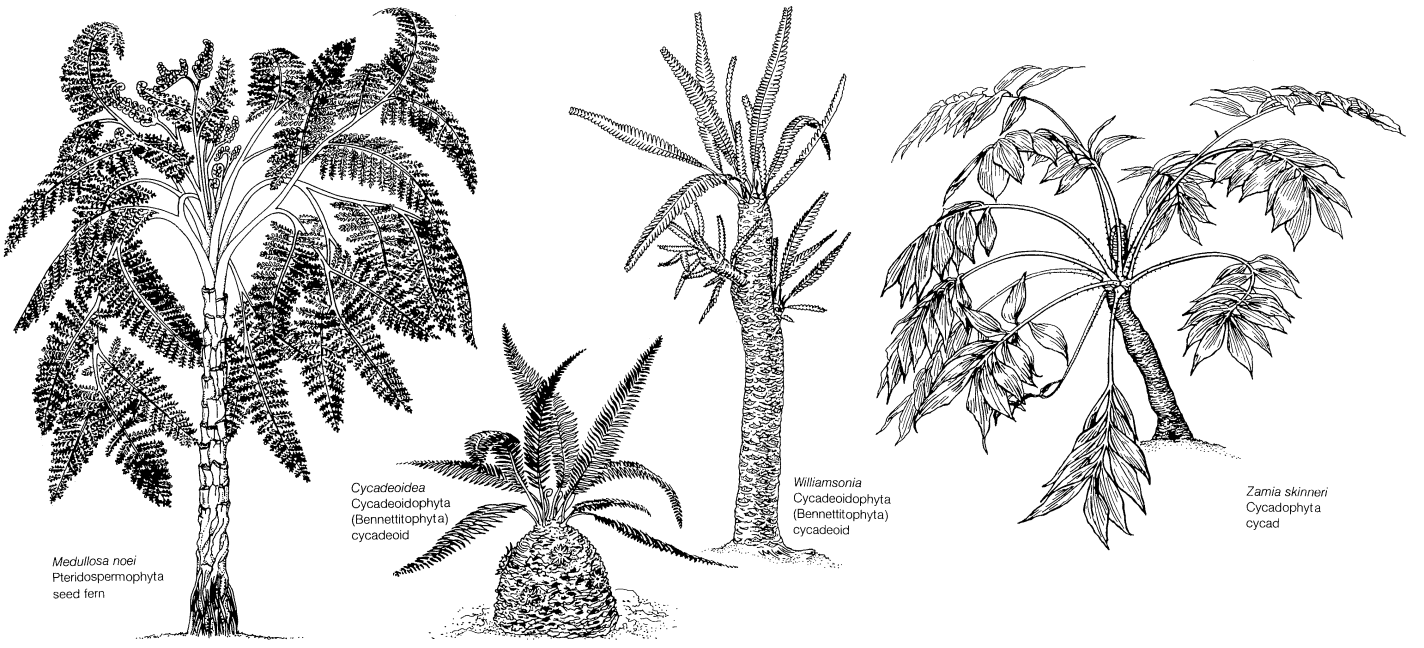


Figure 6: Representative cycadophytes.

From (*Medullosa noei, Cycadeoidea,* and *Williamsonia*) A.S. Foster and E.M. Gifford, Jr., *Comparative Morphology of Vascular Plants,* 2nd ed. (1974), copyright © 1959 and 1974 by W.H. Freeman and Co., *Medullosa noei* redrawn from W.N. Stewart and T. Delevoryas, "The Medullosan Pteridosperms," *Botanical Review,* 22:45 (1956), and *Williamsonia* adapted from C.A. Arnold, *An Introduction to Paleobotany,* copyright © 1947 by McGraw-Hill, Inc.; (*Zamia skinneri*) *Fairchild Tropical Garden Bulletin,* July 1987

In time, many of these "ferns" were recognized as seed plants, and it has been determined that seed ferns were a dominant vegetation in the late Paleozoic. Seed ferns generally are characterized as having been slender trees or, in some cases, woody, climbing vines, but generally with large, fernlike fronds.

Seed-fern foliage

Characteristic seed-fern foliage consisted of large compound leaves composed of second- and sometimes third-order branches (Figure 6). The latter bore fernlike leaflets, hence the name seed fern, although they are only remotely related to true ferns. Seed-fern stems generally possessed variable amounts of soft, loose wood and relatively large zones of cortex and pith; in this respect they resembled the stems of cycads and differed considerably from the stems of conifers, which have compact wood and relatively small zones of cortex and pith.

Reproductive organs of seed ferns were borne upon the foliage; single ovules and seeds were borne in place of pinnae, while male organs often occurred as compound pollen organs composed of partially or wholly united microsporangia. As in other gymnosperms, the ovule consisted of one megasporangium within a single integument. It is believed that, as the reproductive cycle progressed, the megasporangium, also called the nucellus, probably gave rise first to a quartet of megaspores. One of these then produced a large fleshy female gametophyte bearing several archegonia, each with a single egg. Following pollination and fertilization, the ovule developed into a seed with an embryo nested in the fleshy female gametophyte, which served as a food source during germination and seedling growth.

Cycadeoids. Although a few groups of pteridosperms persisted from the late Paleozoic era well into the Mesozoic, the common cycadophytes of the latter ages were members of the Cycadeoidophyta (also known as Bennettitophyta). They are well represented in the later Mesozoic era, well into the Cretaceous period (144 to 66.4 million years ago), by members of the genus *Cycadeoidea,* which had rather squat, barrel-shaped, unbranched trunks and once-pinnate compound leaves (Figure 6). The stems were armoured with the persistent bases of leaves; internally there was a thick pith surrounded by a narrow zone of vascular tissue from which vascular strands extended directly into the leaf bases. The fossilized trunks of these plants display scattered strobili among leaf bases of the characteristic armour. Fossil cycadeoids are widespread but are especially abundant in the Black Hills region of South Dakota.

Earlier in the Mesozoic era, cycadeoids of a more slender, branching form, exemplified by *Williamsonia,* were abundant (Figure 6). As in *Cycadeoidea,* the fronds were single pinnate compound leaves.

Cycadeoid features

The feature that set the cycadeoids apart from other cycadophytes was the compound strobili, which some, but not all, possessed. These strobili were composed of both male and female sporophylls, in some cases subtended by a system of bracts. Although often described as flowerlike and indeed sometimes depicted as having a floral, rosette form, cycadeoid "flowers," unlike true flowers (found in the angiosperms), were composed of sporophylls bearing "naked" (*i.e.,* gymnospermous) ovules. They are not now considered to have given rise to any group of the true angiospermous flowering plants.

Although cycadeoids flourished for millions of years, and must therefore be considered as a highly successful line of plants, they eventually became extinct in the Cretaceous period.

Cycads. The living cycads are for the most part palmlike, cone-bearing plants, generally of low stature (Figure 6). Although few genera, species, and individuals exist, they are extremely important plants in terms of the information that can be gained from studying them. Their reproduction is very primitive in that they rely on flagellated, motile male gametes (spermatozoids), a feature linking them with other plants fertilized by motile flagellated sperm (zooidogamous), such as ferns, club mosses, and other vascular cryptogams. Without knowledge of fertilization in the cycads and *Ginkgo,* it is highly unlikely that scientists would have more than remote theories as to the reproductive modes of seed ferns and other extinct groups of seed plants. Research on cycad reproduction is also providing information on the early origins of insect pollination, long thought to have evolved along with the relatively more recent angiosperms, or flowering plants.

Fossil remains

Distribution and abundance. Seed-fern fossils are found in both the Northern and Southern hemispheres, but many more have been described from Europe and North America than from other regions, primarily because many of the paleobotanical studies are concentrated there. Pteridosperms have been identified in Australia and India in recent years. In both hemispheres, seed ferns are common in coal measures, from which it may be inferred that, ecologically, they were plants of warm humid climates.

Abundant fossils of cycadeoids and cycads have been discovered and described from the Mesozoic era. The oldest remains of undisputed cycads date from the Triassic

period, 245 to 208 million years ago (*e.g., Leptocycas, Antarcticycas*), but some problematic forms (*e.g., Primocycas, Archaeocycas*) are of Paleozoic age. Most Mesozoic cycads resembled extant genera (*e.g., Cycadites, Pseudocycas, Cycadospadix*), and some are referred to present genera (*e.g., Macrozamia zamoides, Zamia coloradensis*). Fossil forms have been found in many places where they are now extinct (for example, Greenland, Antarctica, Alaska, Argentina, France, Austria), testifying to much milder climates in now temperate and even subarctic regions.

Ten genera of cycads are widely recognized. There are three endemic Australian genera—*Macrozamia* (14 species), *Lepidozamia* (two species), and *Bowenia* (two species); four American and Caribbean genera—*Microcycas* (one species), *Zamia* (about 35 species), and *Ceratozamia* and *Dioon* (10 species each); and two African genera—*Encephalartos* (about 40 species) and *Stangeria* (one species). The genus *Cycas,* with about 24 species, is the most wide-ranging, extending from eastern Australia westward across the Pacific and Indian oceans to Madagascar and the east coast of South Africa. In addition to the above well-known genera, a recent collection of cycad specimens from northwestern Colombia included a new genus now described under the name *Chigua. Chigua* reveals features hitherto undescribed in any American genus or species, for the specimens, which in most respects resemble *Zamia,* are unique in having leaflets with midribs and lateral veins, a characteristic formerly known only in *Stangeria.*

Ecology and habitats. Cycads are plants of subtropical habitats, where they occupy a variety of ecological situations ranging from rain forests, to mesophytic savannas, to near-desert scrublands. Now nowhere abundant in nature, wild populations of cycads in many regions are endangered. Only recently have Australian cycads been removed from the noxious weeds list (because of certain toxic properties dangerous to cattle) to a protected status.

NATURAL HISTORY

"Functional conifers"

Sporophyte phase. As in other gymnosperms, the large, woody plant is the sporophyte phase of the life cycle and typically is diploid in chromosome number. All cycads may be called "functional conifers," for all species bear strobili; these strobili are of a simple type, unlike those of true conifers, which bear more complex, compound strobili. It is not considered that this feature of cycads indicates anything other than a parallelism in evolution.

Cycad males and females are morphologically alike except for their sporophylls. Male sporophylls (microsporophylls) are spatulate organs bearing large pollen sacs (microsporangia) in clusters (sori) on their lower (abaxial) surfaces (Figure 7). Up to 200 cubic centimetres of pollen are produced by a single cone of *Cycas rumphii,* and some other species produce similar volumes. It was once estimated that one pollen cone of *Encephalartos* produced seven billion pollen grains having a total volume of about 300 cubic centimetres. While this enormous production would seem to be consistent with a system of wind dispersal, observations and controlled experiments strongly suggest that in most, or perhaps all, cycads, insect pollen vectors are necessary for effective pollination of ovules. The Mexican cycad *Zamia furfuracea,* for example, is pollinated by a small snout weevil, *Rhopalotria mollis,* which lays its eggs and completes its reproductive cycle in male cones. Emerging adults then carry pollen to female cones and pollination of ovules and subsequent fertilization of eggs occurs.

(A) Adapted from A. Engler and K. von Prantl, *Die naturlichen Pflanzenfamilien,* Wilhelm Engelmann, Leipzig, 1916; (B) from A.S. Foster and E.M. Gifford, Jr., *Comparative Morphology of Vascular Plants,* 2nd ed. (1974), copyright © 1959 and 1974 by W.H. Freeman and Co., *Ceratozamia mexicana* redrawn from A. Engler, *Syllabus der Pflanzenfamilien,* vol. 1, copyright © 1954 by Gebruder Borntraeger, West Berlin

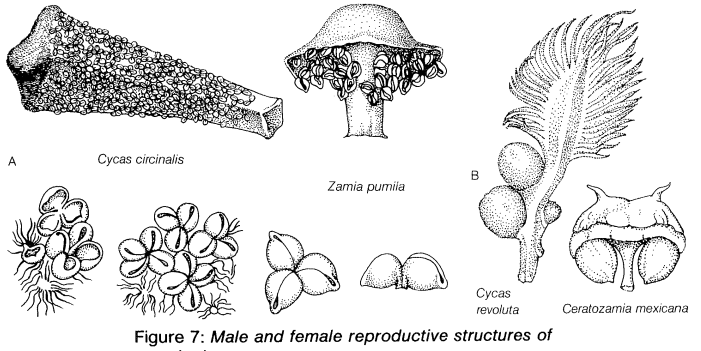


Figure 7: *Male and female reproductive structures of cycadophytes.*
(A) Microsporophylls and microsporangia of two cycads.
(B) Megasporophylls and ovules of two cycads.

Gametophytes

Gametophyte phase. As in all other gymnosperms, male and female sporophytes of cycads produce, respectively, male and female gametophytes. The male gametophyte phase of the life cycle begins in the microsporangium with meiotic production of tetrads of microspores followed by the division of each haploid microspore into a three-celled pollen grain. Because it is multicellular, the pollen grain is considered to be an immature male gametophyte, but its further development into a sexually mature organism occurs only after it has been shed from the microsporangium and transported as a pollen grain to a megasporophyll—specifically, to an ovule, within which the male gametophyte grows to maturity.

Concurrently with pollen development, the ovule differentiates, and at the time of pollination it consists of a large megasporangium (nucellus) enclosed within a fleshy integument. At this time an opening at its distal end (the micropyle) permits pollen to enter the ovule. Over the next three to five months, the male gametophyte develops into a haustorial pollen tube, which eventually penetrates the nucellus and partially projects into an archegonial chamber.

Meanwhile in the nucellus, a single megaspore mother cell undergoes meiosis, forming a tetrad of haploid megaspores, only one of which survives to divide mitotically many times and form a large fleshy female gametophyte. The female gametophyte grows at the expense of nucellar tissue but remains enclosed within its remains. At its micropylar end, this gametophyte develops from one to many archegonia (commonly one to six in most cycads and up to 100 in *Microcycas,* only five or six of which are functional). Each archegonium is composed of a quartet of neck cells beneath which is a large egg. This egg is the largest known in the plant kingdom, being about three millimetres in length.

The development of male and female gametophytes is synchronized, and during the final week or so before fertilization, the male gametophyte forms large, multiflagellated spermatozoids. In all species but *Microcycas,* there is just one pair of sperm per pollen tube; in *Microcycas,* spermatozoids number about 12 to 16 per tube. What actually triggers sperm release into the archegonial chamber is unknown, but when it happens, the spermatozoids quickly move through the archegonial neck into the egg cytoplasm. One sperm loses its flagellature, and fusion of egg and sperm nuclei takes place. Subsequently, the zygote forms a single large embryo, other eggs meanwhile aborting. In the Florida cycad, *Zamia pumila,* the reproductive cycle occurs over a period of about 14 months, cones first becoming visible in October, pollination occurring in December, fertilization taking place in late May and early June, and embryogenesis and seed maturity being completed the following December. Similarly slow reproduction is typical also for other genera and species.

Seed development

As far as is known, cycad seeds have no dormant period or after-ripening requirements and in some cases actually begin germinating while still attached to sporophylls. Possibly some germination inhibitors are present in the outer fleshy layer of the seed, because its removal often accelerates germination, and treatment with scarifying agents also may enhance germinability. The germinating embryo remains attached to the female gametophyte for as long as two years, absorbing nutrition through its cotyledons, which remain embedded in the female gametophyte. The seedling rapidly develops a flesh taproot and root growth.

The outer layer of the seeds of *Cycas circinalis* and *C. rumphii* are thick and somewhat fibrous, and experiments which show them to be capable of long immersion in brine suggest that long-distance dispersal by ocean currents

may account for the presence of these species on remote Pacific islands. Little is known of natural seed dispersal of other cycads, but their bright red seed coats suggest a visual signal to animals. Mockingbirds, squirrels, and coatis are reported to disperse them. In general, however, cycad seeds, which are rather heavy, are poorly dispersed, and most germinate at the base of the parent, where they often languish and die.

FORM AND FUNCTION

Stem. Stems of cycads are characteristically short and stout, and while most genera have some species with subterranean, tuberlike stems, a majority of species are arborescent (Figure 6). The taller cycads include *Microcycas calocoma* (up to 10 metres high), *Macrozamia moorei* (up to 18 metres), *Dioon spinulosum* (up to 16 metres), *Lepidozamia hopei* (up to 18 metres), and *Encephalartos altensteinii* (up to 20 metres), but most of the arborescent (treelike) species have trunks only two to three metres high. The stems of most arborescent species are covered with an armour composed of the hardened leaf and cataphyll bases, but internally they are rather soft and fleshy, with a thick parenchymatous cortex, a large pith, and scanty woody tissue (Figure 8). In most cycads, the woody tissue is on the order of five to 10 millimetres wide, but *Dioon spinulosum* has an exceptional amount of wood, in some specimens up to 10 centimetres wide. This may constitute evidence of the primitive nature of the genus, because seed ferns also generally had stems with considerably more wood than those of most living cycads. Even in *Dioon,* there is no evidence of annual growth rings, so that age estimates must rely on other evidence, most often on counts of the whorls of leaf scars, which can be related to annual or biennial production of new leaf flushes. On this basis, it has been estimated that some cycads (notably *Dioon* and *Macrozamia*) may be as much as 1,000 years old; however, it is doubtful that most cycads are that old.

Stem size

Adapted from (A,B,D) C.J. Chamberlain, *Gymnosperms: Structure and Evolution,* © 1935 by the University of Chicago Press, (B) after Dr. F. Grace Smith; (D) D.W. Stevenson, "Radial Growth in the Cycadales," *American Journal of Botany,* 67:468 (1980)

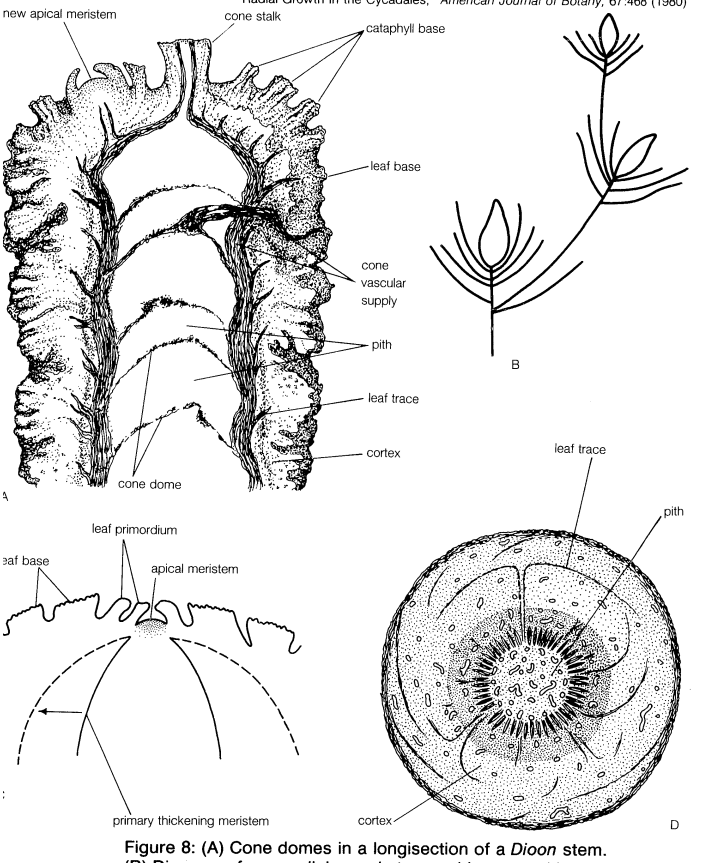


Figure 8: (A) Cone domes in a longisection of a *Dioon* stem. (B) Diagram of sympodial cycad stem architecture with axes much lengthened. (C) Diagram of a five-year-old stem apex of *Zamia,* illustrating lateral growth (arrow) resulting from action of a primary thickening meristem. (D) Cross section of a stem of *Zamia* showing girdling leaf traces.

Species of *Macrozamia, Encephalartos,* and *Cycas* often develop additional cylinders of vascular tissue, apparently formed from vascular cambia originating in the cortex. The result is a condition in which concentric rings of xylem and phloem are present, often two or three, but in exceptional cases, as many as 14. The xylem of cycad seedlings and that of some subterranean stems (*Stangeria, Zamia*) is composed of scalariform tracheids; in older stems, the tracheids exhibit primitive, multiseriate, bordered pits.

Another feature of those cycad stems in which terminal cones are produced is the presence of "cone domes" in the pith (Figure 8A). In longitudinal sections, the pith appears partitioned horizontally at intervals by vascular tissue. Each cone dome represents the displacement of a cone axis to one side as a result of the initiation and growth of the new vegetative apex.

Stem growth

The cycad stem grows from the tip (apically); the only lateral buds and branches are those unusually placed (adventitious) stems, whose buds arise by regeneration after the apical growth tissue (meristem) has been destroyed or as a result of wounding. Apical dominance and lack of branching bring about an apparent single-stemmed (monopodial) growth form, so that older plants become quite palmlike. This appearance, however, is deceptive, because in more than half the genera the apical meristem is converted from a vegetative to a reproductive function in that it is transformed into a strobilus (cone). A new vegetative meristem arises to one side of the cone meristem; subsequent growth and enlargement further displace the cone or cones to the side, so that the monopodial appearance is maintained even though the type of growth is actually sympodial (Figure 8). Only members of three genera (*Macrozamia, Lepidozamia, Encephalartos*) have cones initiated to the side and are truly monopodial; the remaining eight are considered sympodial.

Cycads have such thick stems that rearrangements of internal vascular connectives are not externally apparent. The cycad trunk is about as thick at its crown as at its base, thus furthering the resemblance to palms. Such stems, termed pachycaulous, result as in palms from activity of a primary thickening meristem (PTM) lateral to the apical meristem (Figure 8), which produces much greater increments of cortical parenchyma than would result if only an apical meristem were present. This is an important difference between cycadophytes and coniferophytes, for in the latter there is no PTM and the stem at its apical end is relatively smaller than at its base.

A further characteristic of cycad stems not occurring in cycadeoids, seed ferns, or coniferophytes is the presence of girdling leaf traces. In cacad stems, the vascular strands follow a circuitous route to the leaf bases, which is clearly seen in cross sections of stems. Girdling leaf traces are an important means of distinguishing between cycad and cycadeoid fossils (Figure 8).

Root development

Roots. Cycad seedlings initially form a stout, fleshy taproot that persists in subterranean forms for many years but is augmented by secondary roots which also are quite thick and fleshy. The taproots, larger secondary roots, and, in some cases, underground stems, have contractile elements in the pith and cortex that draw the stem more deeply into the ground.

Branch roots are of two kinds: long-branching geotropic roots and short-branching apogeotropic roots, which are referred to as coralloid because of their irregular, beady appearance. The coralloid roots contain symbiotic cyanobacteria (blue-green algae), which fix nitrogen and, in association with root tissues, produce such beneficial amino acids as asparagine and citrulline.

The taproot does not persist long in arborescent cycads but is replaced by large adventitious roots, which obscure the basic taproot system of the seedling. In all cycads, young roots are diarch with a parenchymatous cortex and an outer cover of epidermal scales. In this aspect they also resemble seed ferns. Older roots become triarch or tetrarch, eventually developing substantial amounts of wood and an outer covering of periderm.

Variation in leaf structure

Leaves. The leaves of cycads are for the most part once-pinnately compound; however, in the genus *Bowenia,* the

leaves are bipinnate and quite fernlike. *Stangeria* also has fernlike leaves, and before cones were found to be associated with them the plant was described as a fern in the genus *Lomaria*. *Stangeria* leaves and those of the recently described *Chigua* are unique in possessing pinnately veined leaflets with midribs and side veins. *Cycas* pinnae also have midribs, but these lack side veins altogether. Pinnae of all other cycads have dichotomously branching, more or less parallel veins. The size of the cycad leaf is variable; *Zamia pygmaea,* the smallest cycad, has leaves about 20–30 centimetres long, while some species of *Macrozamia, Lepidozamia, Ceratozamia,* and *Cycas* have leaves three metres in length.

In cross section, the pinnae of most cycads are rather thick and sclerophyllous. The stomata are sunken and are of the type known as haplocheilic; that is, the guard cells arise directly from the mother cell, as contrasted with the syndetocheilic type, in which the guard cells are one division removed from the mother cell. The haplocheilic type is found in living conifers, pteridosperms, cycads, *Ginkgo,* and some others but not in the *Cycadeoidea.*

Sporophylls and strobili. Cycads are universally dioecious. Male plants produce pollen by leaf homologues called microsporophylls, and female plants produce ovules by leaf homologues known as megasporophylls (Figure 7). In all cycads, the microsporophylls are arranged spirally about a cone axis; in all cycads but *Cycas,* megasporophylls are similarly arranged. Megasporophylls of *Cycas* do not form a true cone but are arranged in two to three whorls at the stem apex. Later the stem resumes vegetative growth, and the megasporophylls then are interposed between whorls of foliar leaves and cataphylls; the usual arrangement is two to three whorls of leaves, then several whorls of cataphylls, followed by megasporophylls, but variations in this sequence are not unusual.

The megasporophyll of the Asian *Cycas revoluta* is considered to most typify the ancestral seed-fern condition. Each megasporophyll consists of a stalk, a fertile portion bearing two to six ovules, and an expanded terminal blade having fringelike "pinnae" (Figure 7). An evolutionary series of plant forms probably led toward the biovulate, peltate megasporophylls of such forms as *Encephalartos, Ceratozamia, Microcycas,* and *Zamia* (Figure 7). Microsporophylls similarly vary among cycads; those of *Cycas* are the more leaflike, those of *Zamia* less so (Figure 7). Microsporangia, which are found on the abaxial surface of microsporophylls, are usually numerous—several hundred in *Cycas,* several dozen in *Zamia*—and arranged in small clusters of two to five. They are the equivalent of sori of ferns and of pteridosperms. The cycad microsporangium resembles a clamshell, being somewhat flattened with an elongate suture (Figure 7).

CLASSIFICATION

Many botanists believe that extant gymnosperms represent at least two evolutionary lineages: one that leads to the extant conifers, taxads, and possibly *Ginkgo* and the Gnetales; and another that leads to the cycadophytes, represented today by seed ferns, cycadeoids, cycads, and perhaps others. Cycadophytes probably had their origins among the Devonian progymnosperms (Progymnospermopsida), although the particular antecedents are unknown.

DIVISION CYCADOPHYTA
Gymnospermous plants possessing compound leaves, ovules have 1 integument; seeds borne on either the foliage or megasporophylls; soft, loose wood contains scalariform tracheids and tracheids with multiseriate bordered pits; stem cross sections show wide zones of pith and cortex.

†DIVISION PTERIDOSPERMOPHYTA (seed ferns)
Primitive, primarily Paleozoic, primarily small trees or woody vines; large compound fronds; leaf-borne ovules; and microsporangia more or less united as synangia; subdivided into several groups; two orders, Caytoniales and Glossopteridales, persisted into Cretaceous, the latter sometimes included with pteridosperms, but commonly ranked separately and thought to be closely related to certain primitive angiosperms.

†DIVISION CYCADEOIDOPHYTA (BENNETTITOPHYTA)
Mesozoic; common and cycadlike; differ from cycads in having direct leaf traces, in sometimes being monoecious, and sometimes having bisexual cones.

DIVISION CYCADOPHYTA
Late Paleozoic? to the present; woody, coniferous plants with compound leaves, simple cones; flagellate motile male gametes; stout, fleshy stems; 4 families currently are recognized.

Family Cycadaceae
Generally restricted to species of *Cycas;* foliar, multiovulate megasporophylls arranged in an indeterminate strobilus; pinnae with a single midrib but lacking lateral, branch veins; 24 species defined.

Family Zamiaceae
Singly pinnate compound leaves, bearing leaflets with parallel, dichotomously branching veins (*Chigua,* if included, would be an exception); simple cones; female cones with biovulate megasporophylls; a total of about 112 species includes *Macrozamia, Lepidozamia, Ceratozamia, Encephalartos, Zamia, Microcycas,* and *Dioon.*

Family Stangeriaceae
Fernlike leaves bearing pinnae with a prominent midrib and numerous dichotomously branching lateral veins; simple cones; female cones with biovulate megasporophylls; includes only *Stangeria paradoxa,* a southern African cycad.

Family Boweniaceae
Differ from other cycads in possessing bicompound leaves; one genus, *Bowenia,* with 2 species. (K.J.No.)

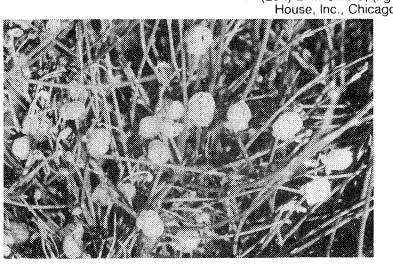
Gnetophytes

The gnetophytes (division Gnetophyta) are characterized as a small group of vascular seed plants that are represented by only three genera: *Ephedra, Gnetum,* and *Welwitschia* (Figure 9). There are about 35 species in the genus *Ephedra,* 30 or more in *Gnetum,* but only one in *Welwitschia.* The three genera exhibit their great diversity in the immense variety of form and size among the various species.

GENERAL FEATURES

Most species of *Ephedra* are branched shrubs or small trees while others are vinelike, often clambering over other vegetation. Species are distributed in dry and cool regions in both the Eastern and Western hemispheres. In the Western Hemisphere, *Ephedra* occurs in desert areas in the southwestern United States, part of Mexico, and a wide area in South America. In the Eastern Hemisphere, certain species occur in China and in the Himalayas from Afghanistan to Nepal and Tibet at elevations of 2,700 to 4,350 metres (10,300 to 14,300 feet). *Ephedra,* known as

(Left) E.M. Gifford; (right) General Biological Supply House, Inc., Chicago; (bottom) Dr. C.H. Bornman



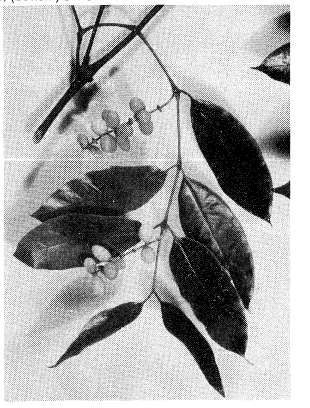
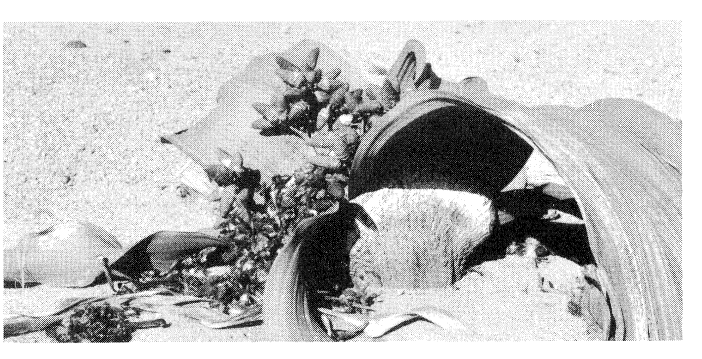


Figure 9: *Representative gnetophytes.*
(Left) Close-up of the shrub and the mature, fleshy seed cones of *Ephedra.* (Right) Foliage leaves and mature seeds of *Gnetum gnemon.* (Bottom) *Welwitschia mirabilis,* showing two gigantic leaves and terminal seed cones on a system of branches.



Importance

Ma-huang, has been a common medicine in China for thousands of years. The effective product, ephedrine, is prescribed for colds, to break a fever and induce sweating, and as a decongestant. Stem fragments of species in the southwestern United States and Mexico are used in the preparation of Mormon tea, Mexican tea, squaw tea, and desert tea. The drug ephedrine is now manufactured synthetically.

Most of the 30 species of *Gnetum* are lianas that climb high into trees of tropical rain forests in central Africa, Asia, northern South America, and islands between Australia and Asia. One species, *G. gnemon,* is a tree about nine metres tall. The leaves are large, much like those of many flowering plants. The seeds are eaten cooked or roasted, and young leaves also are eaten.

The most unusual and geographically restricted gnetophyte is *Welwitschia mirabilis,* which is unlike any other known plant in the world. It occurs in the Namib Desert of southwestern Africa near the coast of Angola and Namibia, as well as inland to about 150 kilometres. (Rainfall on the Namib Desert ranges from zero to 100 millimetres [four inches] per year). There are only two large, permanent, arching leaves; they split, and the tips die when they touch the hot sand. It is not clear how the plant obtains sufficient water to meet its needs. There is a taproot that may extend downward for one to 1.5 metres before it divides into numerous thin roots, which must tap a supply of water not available to other plants of the desert.

FORM AND FUNCTION

Stem. In *Ephedra* the stems are jointed, the basis for the common name joint fir. The leaves are arranged in pairs on the stem or in whorls of three with their bases forming a sheath around the stem at a node. The leaves of most species are reduced or scalelike, although in some species they may grow to three centimetres in length. The leaves of most species have two veins that are connected to two axial stem vascular bundles. The bulk of photosynthesis occurs in the green stems. Reduction in leaf surface may be related to the necessity of reducing water loss through the process of transpiration, a requirement for existence in a desert environment.

Leaves. The leaves of *Ephedra, Gnetum,* and *Welwitschia* are strikingly different in form and venation and provide morphological characters that are definitive for each of the genera.

The leaves of *Gnetum* resemble those of the angiosperms (the flowering plants) in form, structure, and venation. Two leaves at a node are broad and have a pinnate venation system (one midvein with lateral secondary veins that run to the leaf margin) and a meshwork of smaller veins. Older stems become hard by the production of wood (secondary xylem).

Welwitschia

A conspicuous vegetative feature of *Welwitschia* is the presence of two large permanent leaves. When a seed germinates, two seed leaves (cotyledons) emerge, followed by the production of the two permanent leaves. Soon after the development of the two leaves, growth activity shifts away from the tip of the stem (the apical meristem) to the bases of the two permanent leaves. Growth in this region takes place in a meristematic zone, which adds tissue at the base of each leaf. This development results in a bilobed crown and later to a circular concave disk surmounted by a band of meristematic tissue, which continues to contribute new tissue to the two large leaves. Thus, growth in *Welwitschia* has shifted from developing height to developing the two leaves outward. The leaves are perpetuated by the basal meristem at the rate of eight to 15 centimetres per year. The leaves become split and frayed in old plants. A leaf of one giant was reported to have an unbroken width of 1.8 metres and a length of 6.2 metres, of which 3.7 metres were of living tissue. Some plants are estimated to be 1,500 to 2,000 years old.

It is not known for certain how a plant with gigantic leaves such as *Welwitschia* can exist in a desert and carry on photosynthesis if the pores (stomata) of a leaf remain open during the heat of the day. It may be that *Welwitschia* has the type of photosynthesis in which stomata are open only at night. Carbon dioxide (CO_2) is stored in organic acids and then used during the day in photosynthesis. There are also about 100 days of brief morning fog. Water accumulates on the leaves and is thought to enter them, although the process has never been proved. More recently CO_2 has been shown to be taken up during the daylight hours through open stomata, leading to a tremendous loss of water by transpiration for the purpose of cooling the leaf surface.

One of the physical features that distinguish the gnetophytes from other gymnospermous divisions is the presence of vessels in the xylem (wood). A vessel is a longitudinal row of cells, called vessel members, which have several to many circular perforations in their end walls at maturity, providing an efficient pathway for the movement of water in the plant body. The possession of vessels is characteristic of the flowering plants (angiosperms) as well, and has led to speculation that gnetophytes, especially *Gnetum,* may have been close to the ancestral stock of some angiosperms.

Reproductive structures and function. The reproductive structures of gnetophytes are contained within strobili, or cones (Figure 10). Most species are dioecious; *i.e.,* one individual plant produces either pollen-producing or seed-producing cones, called microsporangiate and megasporangiate strobili, respectively. The two types of strobili are basically the same, consisting of oppositely arranged bracts in the axils of which are short fertile shoots.

Strobili and sporophylls

(Left) E.M. Gifford; (right) from A.S. Foster and E.M. Gifford, Jr., *Morphology and Evolution of Vascular Plants,* copyright © 1989 by W.H. Freeman and Co., redrawn from Sanwal, *Phytomorphology,* 12:243 (1962)

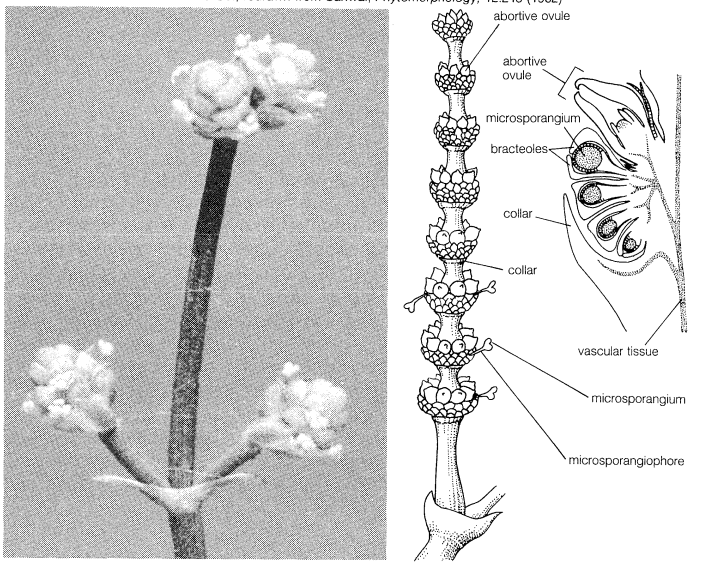


Figure 10: *Reproductive structures of the gnetophytes.* (Left) Pollen-producing strobili (cones) of *Ephedra.* The microsporangiophores extend out between the cone scales, and a pair of scalelike leaves appears at the nodes. (Right) Organization of the microsporangiate strobilus (cone) of *Gnetum gnemon* with (right) a longitudinal section through a node, showing four developing microsporangiate fertile shoots and one abortive ovule.

In *Ephedra* each fertile shoot of a pollen cone consists of an elongated modified structure, a microsporophyll or a microsporangiophore, which terminates in a cluster of sporangia, called microsporangia, that house the haploid reproductive cells (spores). The microsporangia are surrounded by a pair of bracteoles (scalelike leaves). Meiotic divisions in cells of the microsporangia produce the haploid pollen grains.

When shed from the microsporangia, the pollen grains (or male gametophytes) consist of five cells contained within the pollen-grain wall. The pollen grains are boat-shaped, with longitudinal ridges and furrows that run from one end of the grain to the other. Pollen of this type has been identified in the fossil record from the Permian (286 to 245 million years ago) to the present.

Megasporangiate, or seed-producing, strobili (female cones) consist of oppositely paired bracts in the axils of which are fertile shoots consisting of paired bracteoles

enclosing an ovule—the forerunner of a seed. The ovule consists of a delicate inner envelope, called an integument, that encloses a tissue (nucellus) in which a cell divides meiotically to produce a row of haploid cells called megaspores. One megaspore greatly enlarges and undergoes mitotic divisions, producing multiple nuclei that are not surrounded by walls. After 500 to 1,000 nuclei are produced, cell walls begin to form, converting the megagametophyte (or female gametophyte) into a cellular structure. At the upper end, egg-bearing protective structures called archegonia are formed, each of which contains a haploid egg cell.

Pollination and embryogeny. Just before the transfer of pollen grains to the ovule (pollination) some cells of the nucellus degenerate and a viscous solution is extruded through the opening (micropyle) of the now-extended integument. A pollination drop is formed at the tip of the integument. Transfer of pollen may be either by wind or insects. The pollination drop is very high in sugar content and attracts insects. As the sugary pollination drop dries, the pollen grains are pulled inward through the integument and come to rest on the nucellus. The pollen grain germinates, and a pollen tube enters the egg cell of the archegonium. A mature male gamete (sperm) fuses with the mature egg nucleus, bringing about fertilization. Eight or more diploid nuclei are produced and soon become surrounded by cellulose-containing walls. Each of these young embryos may develop, although only one generally reaches a fully developed stage in the mature seed. A seed consists of an embryo with two seed leaves (cotyledons), a stem axis, and a root, embedded in nutritive tissue of the female gametophyte. The pair of protective bracteoles become hard, and the seed is also surrounded by fleshy bracts that may become ivory, red, or orange in colour, perhaps an adaptation for animal dispersal. There appears to be no resting stage or dormant period for the seed, which may germinate immediately upon dispersal.

Fertilization

The strobili of *Gnetum* are compact or may become elongate with conspicuous nodes and internodes. In a microsporangiate, or pollen-producing, strobilus there are two fused bracts at each node forming a cup-shaped structure (collar) that encloses numerous fertile shoots. Each fertile unit consists of two bracteoles enclosing a microsporophyll or microsporangiophore with two sporangia at the tip. There may also be a top whorl of abortive ovules at each node.

The megasporangiate (ovulate) strobilus likewise consists of conspicuous nodes, each with a fused pair of bracts subtending several ovules. Each ovule comprises an integument and two ensheathing structures called envelops. In *Gnetum,* in contrast to *Ephedra,* fertilization occurs before the haploid female gametophyte becomes completely cellular. Nuclei become surrounded by membranes. A pollen tube pushes into the female gametophyte, and a male gamete fertilizes an egg. Several pollen tubes may enter the female gametophyte, resulting in more than one fertilization, although normally only one embryo develops to maturity. The seeds become fleshy and often are brightly coloured.

Reproduction in *Welwitschia*

In *Welwitschia,* branched reproductive shoot systems arise from the meristematic tissue at the bases of the two large permanent leaves. In both types of strobili, paired bracts and fertile shoots develop from the axils, as is true of the other two genera of gnetophytes. The fertile shoot of a microstrobilus has two pairs of bracteoles, the inner pair of which are fused, enclosing six microsporangiophores and a sterile ovule. In a megastrobilus the fertile shoot is made up of a pair of basal bracteoles and a pair of fused bracteoles surrounding an ovule.

At the time of pollination the integument of a fertile ovule in the female elongates and a pollination drop is formed at the tip. A cell in the nucellus of the female ovule undergoes meiosis, but no cell walls are formed between the four haploid nuclei. The four nuclei and their derivatives undergo repeated mitoses without cell-wall formation. Only later do cell walls form around several nuclei. No archegonia are formed. Subsequent development of the female gametophyte is without parallel in other gymnosperms or in angiosperms (Figure 11).

Some of the multinucleate female gametophyte cells grow into the nucellus of the ovule, into which the nuclei migrate. Pollen tubes, which grow downward in the nucellus from the male pollen grains after pollination, meet and fuse with the upwardly growing multinucleate tubes of the female gametophyte. "Fertilization bulbs" are formed, within which fusion takes place between a male gamete and one nucleus of the female gametophyte cells. A zygote is formed, and the young embryo then grows downward within the female gametophyte tube toward the cellular female gametophyte.

The later stages of embryo development take place within the tissue of the female gametophyte. A mature, ripe seed consists of a dicotyledonous embryo embedded in the female gametophyte and surrounded by papery fused bracteoles, constituting a "wing," which probably assists in seed dispersal.

From A.S. Foster and E.M. Gifford, Jr., *Morphology and Evolution of Vascular Plants,* copyright © 1989 by W.H. Freeman and Co.

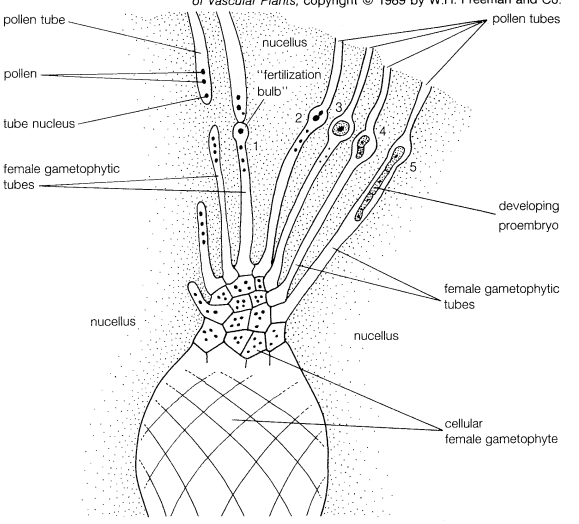


Figure 11: (Left to right) The stages in fertilization and proembryo development in *Welwitschia.* (1) Certain multinucleate female gametophyte cells form tubes that grow up into the nucellus. Pollen tubes growing down in the nucellus meet the upward-growing tubes of the female gametophyte. (2) Wall dissolution occurs, and fertilization takes place in the "fertilization bulb" to form a zygote (3). (4) The zygote divides, forming a two-celled proembryo. (5) The proembryo grows down inside the female gametophytic tube toward the nutritive tissue below.

CLASSIFICATION

Annotated classification.

DIVISION GNETOPHYTA

Longitudinal files of cells with pores in end walls (vessels) in xylem (wood); dioecious; strobili (cones) consist of bracts, in the axils of which are the reproductive structures, which have been likened to flowers of angiosperms.

Order Ephedrales

Shrubs to small trees; small leaves with 2 or 3 veins; mature cones often become fleshy and brightly coloured; 1 family, Ephedraceae; 1 genus, *Ephedra,* with about 35 species.

Order Gnetales

Mostly vines, but a few trees; large flat leaves that have reticulate venation; seeds may be brightly coloured; 1 family, Gnetaceae; 1 genus, *Gnetum,* with about 30 species.

Order Welwitschiales

Monotypic; 2 immense, permanent leaves, which become split and frayed with age; seeds with winglike extensions that may aid in dispersal; restricted to Namib Desert of Africa and vicinity; 1 family, Welwitschiaceae, with 1 member, *Welwitschia mirabilis.*

Critical appraisal. There have been attempts to demonstrate that the gnetophytes form a link between gymnosperms and angiosperms. Proponents of this concept cite the compound nature of both the pollen-producing and seed-producing strobili (some botanists have interpreted the strobili as "inflorescences"; *i.e.,* the axillary fertile shoots of both types of cones are considered to be flowers); and the presence of vessels in the xylem (wood),

a major difference between gnetophytes and all other gymnosperms. In recognition of these similarities, a classification has been proposed that placed the gnetophytes in the group Chlamydospermae, a group between the gymnosperms and the angiosperms. The word "chlamydospermae" means a seed with a "cloak," or envelope, a reference to pairs of bracteoles (scalelike structures) that surround a seed, similar to the ovary wall of the flowers of angiosperms. In another system the three genera are placed in one family in the order Gnetales.

More recently, the trend has been to establish a major group, the Gnetophyta (a division) or Gnetopsida (a class), consisting of three orders—Ephedrales, Gnetales, and Welwitschiales—each comprising one family and one genus. This system recognizes that there are certain shared characteristics but at the same time recognizes that there are significant differences between them.

The existence of gnetophytes in the fossil record has been based primarily upon the presence of pollen grains resembling those of *Ephedra* and *Welwitschia.* Leafy shoots with gnetalean features from the Early Cretaceous have been found associated with *Welwitschia*-type pollen.

Ginkgophytes

Maidenhair tree

Ginkgo biloba is also called maidenhair tree because of the bilobed nature of its leaves and its resemblance to the leaves of the maidenhair fern. *G. biloba* may be the oldest living seed plant, and it is regarded by some as one of the wonders of the world.

GENERAL FEATURES

Fossil leaves with similar form and venation to the living *Ginkgo* have been found in the Jurassic period (208 to 144 million years ago). These fossils have been described from such geographically separated areas as Australia, western North America, Mongolia, Alaska, England, and central Europe. The fossils vary greatly in form and are usually described as species of the genus *Ginkgoites.* Almost the same degree of variation in leaf form can be found on a living *Ginkgo* tree, however. Some paleobotanists, therefore, have recommended the abandonment of the genus *Ginkgoites* and the recognition of several species of *Ginkgo.*

There is one type of ginkgophyte leaf in the fossil record that is generally regarded as a distinct form and is given the generic designation *Baiera.* The leaf is deeply lobed into four segments and lacks a stalk (petiole). Following the Mesozoic era, *Ginkgo* declined progressively in its distribution, and some botanists believe that China was the last natural home of the maidenhair tree. Whether *Ginkgo* still exists in the wild state in China is unsettled. Although some evidence supports the view that the stands of trees in southeastern China are of a natural origin, some botanists contend that they may represent the offspring of trees cultivated in temple gardens for thousands of years. The latter may well be the case because, after the outer fleshy seed coat is removed, the seed kernel has been, and is still today, used as food in China and Japan. For 3,000 years or longer, extracts of the *Ginkgo* leaf have been recommended in Chinese medicine as benefiting the heart and lungs. *Ginkgo* has been investigated for its effects in the treatment of asthma, toxic shock syndrome, and various circulatory disorders.

Distribution of *Ginkgo*

FORM AND FUNCTION

Stem. At maturity a *Ginkgo* tree can reach heights of 20 to 30 metres (65 to 100 feet). Young trees often have a central trunk with regular, lateral branching; in older trees the branching is irregular.

A conspicuous feature of *Ginkgo* is the possession of long branches and short, or spur, branches. Leaves are produced on long branches during the spring growth. In subsequent years, clusters of leaves are formed on the lateral short branches. *Ginkgo* is deciduous, and the leaves of some varieties turn a beautiful golden colour in the autumn. The colour varies to some extent among horticultural varieties. There is some plasticity in growth form in that a short branch may become a long branch or the tip of a long branch may be converted to a short branch. The interplay between these two types of branches accounts for the more irregular shape of the older trees. Branching appears to be controlled by the distribution of auxin, a naturally occurring plant hormone.

The trunk diameters of the older specimens of *Ginkgo* may become large as a result of secondary growth. The vascular cambium gives rise to secondary phloem and secondary xylem (wood) for the conduction of water and dissolved minerals. The growth activity of the vascular cambium is sustained in the trunk and long shoots and produces a rather hard wood with well-defined growth rings. The activity of the vascular cambium persists in short shoots, but only a limited amount of soft wood is produced each year.

Leaves. One of the most distinctive features of *G. biloba* is the foliage leaf (Figure 12), which consists of a leaf stalk (petiole) and a fan-shaped dichotomously veined blade, or lamina. Although *biloba* (or bilobed) correctly describes the form of many *Ginkgo* leaves, there is a great range of variation in the degree of lobing and dissection among leaves of the same tree. Bilobed and undivided leaves occur on spur or short branches, while most of the leaves on the upper part of a long branch are divided by a deep sinus into two lobes, each of which is further dissected into segments. Multilobed leaves also occur on new branches (sucker shoots) arising from the tree trunk at ground level.

Variation in leaf lobing

(Top) Sonia Cook, (bottom) John H. Gerard

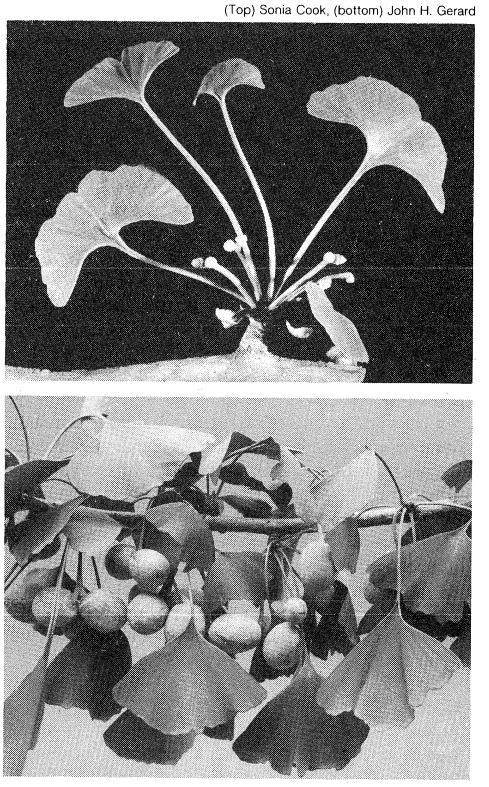


Figure 12: *Representative features of Ginkgo.* (Top) Ovuliferous structures on spur branch. Note bilobed, fan-shaped structure of the leaf. (Bottom) Mature seeds attached to a short branch.

The dichotomous venation pattern in a leaf blade is a striking morphological characteristic of *Ginkgo.* Two vascular bundles extend through the petiole and give rise to two systems of dichotomously branched veins. This type of venation was also present in the leaves of extinct members of the Ginkgoales. Such a system of venation is often referred to as an open type, devoid of vein fusions. It has been shown, however, that vein unions may occur with some regularity.

Reproductive structures and function. Completion of the entire reproductive cycle, from the advent of pollination to the production of seeds with well-developed embryos, takes about 14 months. Pollination and the de-

velopment of the sexual, or gametophytic, phase of the life cycle occur in the first year (April to September), but embryo development is not completed until the spring of the following year.

Ginkgo is dioecious, which means that pollen-producing structures and ovules are produced on separate trees. The reproductive structures are restricted to the spur branches, where they are evident in the spring in the axils of bud scales and foliage leaves.

The pollen-producing strobilus is a loose, pendulous, catkinlike structure consisting of a main axis to which are attached numerous appendages, each of which usually bears two microsporangia at its tip. Meiosis occurs in cells of the microsporangia, giving rise to numerous haploid microspores. Cell divisions take place within the microspores, resulting in the formation of five-celled pollen grains (male gametophytes).

Ovuliferous structures

Ovuliferous structures (Figure 12) also arise in the axils of bud scales and the foliage leaves of spur branches. Each consists of a stalk that bears two or sometimes three or more erect ovules. An ovule is composed of an integument (the future seed coat) surrounding a tissue called the nucellus. It is in the nucellus that meiosis occurs, resulting in the formation of four haploid megaspore cells. It is at about this time that pollen grains are released from the microsporangia of male trees. The pollen (male gametophyte) is carried by wind currents and adheres to a pollination droplet, which exudes from the micropyle at the tip of the integument. Retraction of the droplet brings the pollen grains into a pollen chamber in the nucellus, where they develop into multibranched pollen tubes (male gametophytes).

One of the megaspores in the ovule that results from meiosis enlarges and undergoes a succession of free nuclear divisions (without wall formation). After about 8,000 haploid nuclei are produced, cell walls begin to form. After the female gametophyte becomes cellular, archegonia (normally two) are initiated at the surface toward the micropylar end of the ovule. An archegonium consists of neck cells and a large egg cell.

The basal end of the filament-like male gametophyte becomes suspended in a cavity above the female gametophyte (called the fertilization chamber). The spermatogenous cell of a male gametophyte divides, resulting in the production of two multiflagellated sperm. The sperm and the contents of the pollen tube are released into the fertilization chamber. The sperm swim in the liquid for a brief period of time. Approximately 1,000 flagella are attached to a spiral band at the anterior end. A sperm enters an archegonium and fuses with the egg nucleus. *Ginkgo* and the cycads are the only seed-producing plants that have motile sperm.

Fertilization

The growth of the embryo (embryogenesis) may begin shortly after fertilization but continues after the developing seeds fall to the ground. The embryo grows into the nutritive tissue of the female gametophyte. A seed at maturity (Figure 12) consists of a dicotyledonous embryo, nutritive tissue of the female gametophyte, and the seed coat, which is made up of a hard inner layer and a fleshy, orange-coloured outer layer. Because of the presence of butyric acid, upon decay the fleshy layer emits an odour similar to rancid butter.

CLASSIFICATION

In earlier systems of classification, *Ginkgo* was placed in the class Coniferopsida, along with conifers (*e.g.*, pine, fir, spruce). In recent years *Ginkgo* and its fossil allies have been placed in a separate group, the division Ginkgophyta (sometimes classified as the class Ginkgopsida), in recognition of the many characteristics outlined above. The early ancestral stock of *Ginkgo* extends back in the fossil record to a time coordinate with the ancestors of the conifers, but the two groups appear to have evolved independently.

DIVISION GINKGOPHYTA

Large trees; leaves typically fan-shaped and bilobed, or with more lobes, especially in fossil forms; leaves borne mainly on spur (short) branches; male and female trees; seeds with a fleshy outer layer that upon decay, emits an odour of rancid butter; a single order, Ginkgoales; a single family, Ginkgoaceae; a single extant genus, *Ginkgo;* 2 extinct genera, *Ginkgoites* and *Baiera.* (E.M.G.)

BIBLIOGRAPHY. General works providing comprehensive coverage of the gymnosperms include K.R. SPORNE, *The Morphology of Gymnosperms: The Structure and Evolution of Primitive Seed-Plants,* 2nd ed. (1974), a compact summary discussing both living and extinct groups; THOMAS N. TAYLOR, *Paleobotany: An Introduction to Fossil Plant Biology* (1981), an excellent survey of fossil plants, including the history of the various gymnosperm groups and especially strong on the evolution of seeds; CHARLES JOSEPH CHAMBERLAIN, *Gymnosperms: Structure and Evolution* (1935, reprinted 1966), a classic description of the life history and morphology of all extant groups; W. DALLIMORE and A. BRUCE JACKSON, *A Handbook of Coniferae and Ginkgoaceae,* 4th ed., rev. by S.G. HARRISON (1966), a well-illustrated discussion of many representative types, including cultivated forms; and ERNEST M. GIFFORD and ADRIANCE S. FOSTER, *Morphology and Evolution of Vascular Plants,* 3rd ed. (1989), focusing on the structure and reproduction of vascular plants, including the gymnosperms.

For conifers, see GERD KRÜSSMANN, *Manual of Cultivated Conifers,* 2nd rev. ed. (1985; originally published in German, 1983), an extensively illustrated account of cultivars and species of conifers used in gardens; D.M. VAN GELDEREN and J.R.P. VAN HOEY SMITH, *Conifers* (1986), a pictorial work complementing Krüssmann's book with colour photographs; N.T. MIROV, *The Genus Pinus* (1967), a thorough treatment of all aspects of the biology of the many species of this important temperate genus; GEORGE S. ALLEN and JOHN N. OWENS, *The Life History of Douglas Fir* (1972), a detailed description of the reproductive cycle of the conifer; RUDOLF FLORIN, "Evolution in Cordaites and Conifers," *Acta Horti Bergiani* 15(11):286–388 (1951), a summary of the definitive research on the evolution of seed cones in coniferophytes, and "The Distribution of Conifer and Taxad Genera in Time and Space," *Acta Horti Bergiani* 20(4):122–312 (1963), an important survey, with thorough maps of distributions of living and fossil coniferophytes; CHARLES B. BECK (ed.), *Origin and Evolution of Gymnosperms* (1988), a collection of studies of fossil conifers with important discussions of early species; J.A. HART, "A Cladistic Analysis of Conifers: Preliminary Results," *Journal of the Arnold Arboretum* 68:269–307 (July 1987), a new approach to discovering relationships among conifer genera; and J.E. ECKENWALDER, "Re-evaluation of Cupressaceae and Taxodiaceae: A Proposed Merger," *Madroño* 23(5):237–256 (1976), an original look at family relationships among extant conifers.

A survey of cycadophytes is found in HAROLD C. BOLD, CONSTANTINE J. ALEXOPOULOS, and THEODORE DELEVORYAS, *Morphology of Plants and Fungi,* 5th ed. (1987). Other works include CHARLES JOSEPH CHAMBERLAIN, *The Living Cycads* (1919, reprinted 1965), the best comprehensive work on the cycads; CYNTHIA GIDDY, *Cycads of South Africa,* 2nd rev. ed. (1984), an excellent introduction to cycad morphology, noted for its beautiful colour illustrations of *Encephalartos* species in natural habitats; DIVYA DARSHAN PANT, *Cycas and the Cycadales,* 2nd ed. (1973), a fine presentation of all aspects of the life of *Cycas* and a valuable general reference to their anatomy and morphology; and PÁL GREGUSS, *Xylotomy of the Living Cycads, with a Description of Their Leaves and Epidermis,* trans. from Hungarian (1968), which emphasizes the structure of cycad xylem and includes important information on the habits and leaf morphology of cycads.

General treatments of the gnetophytes and ginkgophytes are found in PETER H. RAVEN, RAY F. EVERT, and SUSAN E. EICHHORN, *Biology of Plants,* 4th ed. (1986). WILSON N. STEWART, *Paleobotany and the Evolution of Plants* (1983), offers a more detailed account of the fossil record of *Ginkgo;* and CHRIS H. BORNMAN, *Welwitschia: Paradox of a Parched Paradise* (1978), is a brief study of *Welwitschia mirabilis.*

(T.De./K.J.No./J.E.Ec./E.M.G.)

The House of Habsburg

The House of Habsburg (also spelled Hapsburg; also known as the House of Austria) was one of the greatest of the sovereign dynasties of Europe.

ORIGINS

The name Habsburg is derived from the castle of Habsburg, or Habichtsburg ("Hawk's Castle"), built in 1020 by Werner, bishop of Strasbourg, and his brother-in-law, Count Radbot, in the Aargau overlooking the Aar River, in what is now Switzerland. Radbot's grandfather, Guntram the Rich, the earliest traceable ancestor of the house, may perhaps be identified with a Count Guntram who rebelled against the German king Otto I in 950. Radbot's son Werner I (died 1096) bore the title count of Habsburg and was the grandfather of Albert III (died *c.* 1200), who was count of Zürich and landgrave of Upper Alsace. Rudolf II of Habsburg (died 1232) acquired Laufenburg and the "Waldstätte" (Schwyz, Uri, Unterwalden, and Lucerne), but on his death his sons Albert IV and Rudolf III partitioned the inheritance. Rudolf III's descendants, however, sold their portion, including Laufenburg, to Albert IV's descendants before dying out in 1408.

AUSTRIA AND THE RISE OF THE HABSBURGS IN GERMANY

Albert IV's son Rudolf IV of Habsburg was elected German king as Rudolf I in 1273. It was he who, in 1282, bestowed Austria and Styria on his two sons Albert (the future German king Albert I) and Rudolf (reckoned as Rudolf II of Austria). From this date the age-long identification of the Habsburgs with Austria begins (see AUSTRIA). The family's custom, however, was to vest the government of its hereditary domains not in individuals but in all male members of the family in common, and, though Rudolf II renounced his share in 1283, difficulties arose again when King Albert I died (1308). After a system of condominium had been tried, Rudolf IV of Austria in 1364 made a compact with his younger brothers that acknowledged the principle of equal rights but secured de facto supremacy for the head of the house. Even so, after his death the brothers Albert III and Leopold III of Austria agreed on a partition (Treaty of Neuberg, 1379): Albert took Austria, Leopold took Styria, Carinthia, and Tirol.

King Albert I's son Rudolf III of Austria had been king of Bohemia from 1306 to 1307, and his brother Frederick I had been German king as Frederick III (in rivalry or conjointly with Louis IV the Bavarian) from 1314 to 1330. Albert V of Austria was in 1438 elected king of Hungary, German king (as Albert II), and king of Bohemia; his only surviving son, Ladislas Posthumus, was also king of Hungary from 1446 (assuming power in 1452) and of Bohemia from 1453. With Ladislas the male descendants of Albert III of Austria died out in 1457. Meanwhile the Styrian line descended from Leopold III had been subdivided into Inner Austrian and Tirolean branches.

Frederick V, senior representative of the Inner Austrian line, was elected German king in 1440 and crowned Holy Roman emperor, as Frederick III, in 1452—the last such emperor to be crowned in Rome. A Habsburg having thus attained the Western world's most exalted secular dignity, a word may be said about the dynasty's major titles. The imperial title at this time was, for practical purposes, hardly more than a glorification of the title of German king; and the German kingship was, like the Bohemian and the Hungarian, elective. If Habsburg was to succeed Habsburg as emperor continuously from Frederick's death in 1493 to Charles VI's accession in 1711, the principal reason was that the hereditary lands of the Habsburgs formed an aggregate large enough and rich enough to enable the dynasty to impose its candidate on the other German electors (the Habsburgs themselves had an electoral vote only in so far as they were kings of Bohemia).

The emperor Frederick III

For the greater part of Frederick's reign it was scarcely foreseeable that his descendants would monopolize the imperial succession so long as they did. The Bohemian and Hungarian kingdoms were lost to the Habsburgs for nearly 70 years from the death of Ladislas Posthumus in 1457; the Swiss territories, lost in reality from 1315 onward (see SWITZERLAND), were finally renounced in 1474; and Frederick's control over the Austrian inheritance itself was long precarious, not only because of aggression from Hungary but also because of dissension between him and his Habsburg kinsmen. Yet Frederick, one of whose earliest acts in his capacity as emperor had been to ratify, in 1453, the Habsburgs' use of the unique title of "archduke of Austria" (first arrogated for them by Rudolf IV in 1358–59), may have had some prescient aspiration toward worldwide empire for the House of Austria: the motto *A.E.I.O.U.*, which he occasionally used, is generally interpreted as meaning *Austriae est imperare orbi universo* ("Austria is destined to rule the world"), or *Alles Erdreich ist Österreich untertan* ("The whole world is subject to Austria"). He lived long enough to see his son Maximilian make the most momentous marriage in European history (see below); and three years before his death he also saw the Austrian hereditary lands reunited when Sigismund of Tirol abdicated in Maximilian's favour (1490).

Before explaining what the Habsburgs owed dynastically to Maximilian, mention can be made of a physical peculiarity characteristic of the House of Habsburg from the emperor Frederick III onward: his jaw and his lower lip were prominent, a feature supposed to have been inherited by him from his mother, the Mazovian princess Cymbarka. Later intermarriage reproduced the "Habsburg lip" more and more markedly, especially among the last Habsburg kings of Spain.

THE WORLD POWER OF THE HABSBURGS

Even before Frederick III's time the House of Habsburg had won much of its standing in Germany and in central Europe through marriages to heiresses. Frederick's son Maximilian carried this matrimonial policy to heights of unequalled brilliance. First he himself in 1477 married the heiress of Burgundy, Charles the Bold's daughter Mary, with the result that the House of Habsburg, in the person of their son Philip, inherited the greater part of Charles the Bold's widespread dominions: not the duchy of Burgundy itself, which the French seized, but Artois, the Netherlands, Luxembourg, and the County of Burgundy or Franche Comté. Secondly, though he failed after Mary's death in 1482 to secure Brittany also by a similar coup (France frustrated his proxy marriage to the Breton heiress Anne), he procured Philip's marriage, in 1496, to Joan, prospective heiress of Castile and Aragon: thus securing for his family not only Spain, with Naples–Sicily and Sardinia, but also the immense dominions the Spaniards were about to conquer in America. Maximilian's matrimonial achievements were the occasion of the famous hexameter *Bella gerant alii, tu felix Austria nube* ("Let others wage wars: you, fortunate Austria, marry").

Maximilian: the Burgundian and Spanish marriages

Since Philip I of Castile died prematurely, his son was already ruler of the Burgundian heritage and of Spain when, in 1519, he succeeded Maximilian as ruler of the Habsburgs' Austrian territories. In the same year he was elected Holy Roman emperor as Charles V.

The threat of force as well as an enormous expenditure in bribes was necessary to secure Charles's election. Besides the fact that many of the German princes were reluctant to saddle themselves with so mighty a sovereign, there was the opposition of France, which saw itself already half-encircled, from the northeast clockwise to the southwest, by Charles's possessions. Dating from Maximilian's Burgundian marriage, antagonism between the French kings

Charles V

and the Habsburgs was to persist, to the progressive detriment of the latter, until the middle of the 18th century, and until the second half of the 17th the other European powers would mostly sympathize with France. The Habsburgs in the 16th century were too formidable not to provoke envy and anxiety.

Charles V's responsibilities at the time of his becoming emperor were moreover too great for one man to assume, as he himself could acknowledge: they had to be divided. By the Treaty of Brussels (1522) he assigned the Habsburg-Austrian hereditary lands to his brother, the future emperor Ferdinand I. In 1521 Ferdinand had married Anna, daughter of Louis II of Hungary and Bohemia; and Louis II's untimely death in 1526, after his defeat by the Turks in the Battle of Mohács, prompted Ferdinand to stand as candidate for his succession, to which, despite rivals, he was elected.

The Habsburgs reached the zenith of their power before the end of the 16th century: the duchy of Milan, annexed by Charles V in 1535, was assigned by him to his son, the future Philip II of Spain, in 1540; Philip II conquered Portugal in 1580; and the Spanish dominions in America were ever expanding. There were, however, three faults in the power structure—two of them historical accidents, the

Table 1: The Imperial Succession of the Habsburgs, 1493–1740

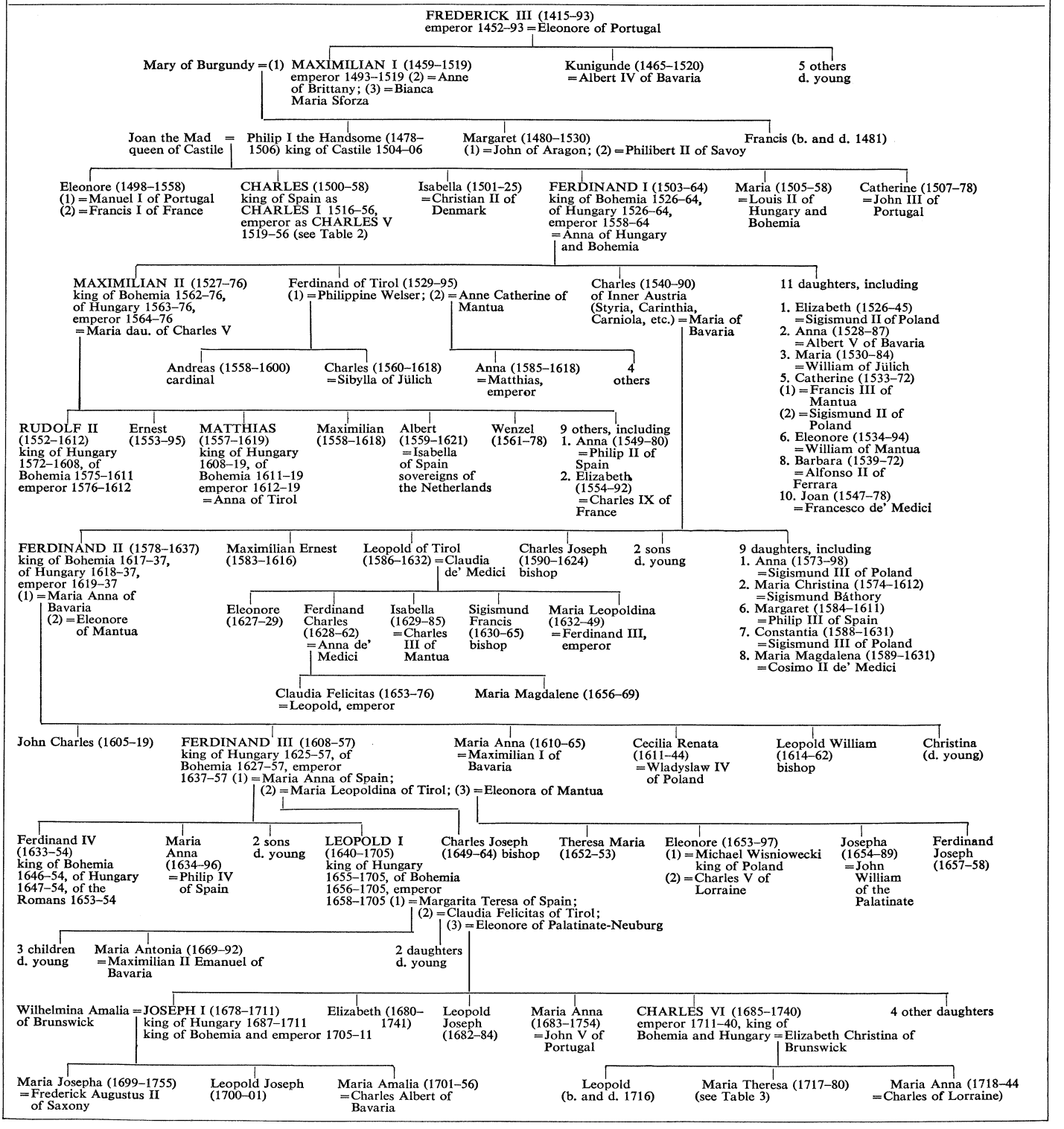
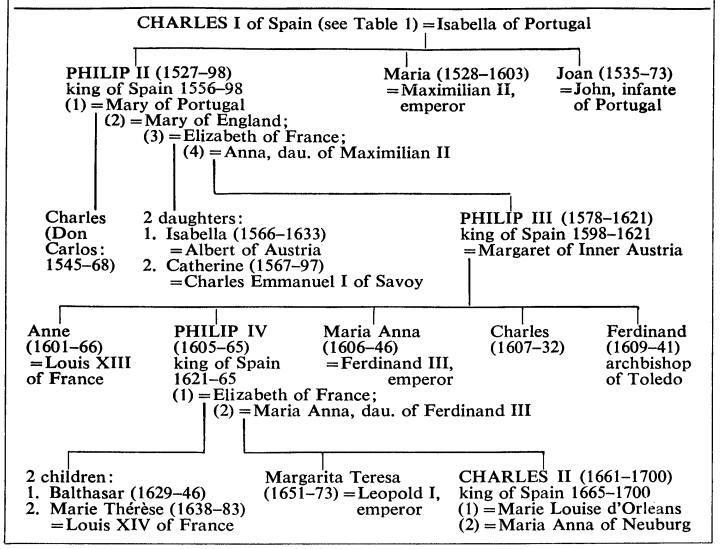


Table 2: The Spanish Habsburgs



External and internal weaknesses in the 16th century

third an effect of the Habsburg dynasty's own measures for self-preservation.

In the first place, the ascendancy of Charles V coincided with the outbreak of the Protestant Reformation in Germany, which was to spread turmoil for decades over Europe from the Netherlands to Hungary. As Charles, from his Spanish upbringing, was imbued with ideas of Catholic uniformity and as his successors, with the exception of the enigmatic Maximilian II, sought also to realize those ideas, religious resistance to the Habsburgs' authority came to aggravate or to camouflage political resistance. At the same time, the papacy, overawed though it was by the Spanish military presence in Italy, did not always subscribe to the Habsburg's special policy for Catholicism.

Secondly, Ferdinand's accession to Hungary meant that the Habsburgs had to bear the brunt of the Ottoman Turkish drive from the Balkans into central Europe, just as Habsburg Spain had to confront Turkish incursions into the western Mediterranean. The great victory of Lepanto (1571), won by a Habsburg bastard, did not end these troubles, which were exploited, against the dynasty, by Hungarian dissidents and, more covertly, by France.

The third flaw in the Habsburg edifice was latent in the 16th century. Mindful of what they had won by marriages, the Habsburgs sought to preclude rival dynasties from turning the tables on them by the same means: to keep their heritage in their own hands, they began to intermarry more and more frequently among themselves. The result, in a few generations, was a fatal inbreeding that brought the male line of Charles V to extinction.

The imperial and the Spanish lines

By a series of abdications toward the end of his life Charles V transferred his Burgundian, Spanish, and Italian possessions to his son Philip II and his functions as emperor to his brother Ferdinand, who succeeded him formally as such after his death (1558). This division of the dynasty between imperial and Spanish lines was definitive: Ferdinand's male descendants were Holy Roman emperors until 1740 (for the emperors from Frederick III to Charles VI see Table 1), Philip's were kings of Spain until 1700 (see Table 2). The imperial line was inevitably concerned to maintain its position in Bohemia and to assert itself against the Turks in divided Hungary, because the loss of the two kingdoms would have meant the reduction of its possessions to what the Habsburgs had had hereditarily before Frederick III's time (the Austrian duchies and scattered holdings in Swabia and in Alsace)—a reduction that in turn would have compromised its chances of continuing to be elected to the German kingship. Philip II of Spain remained territorially the greatest sovereign in the Western world until his death in 1598; but the Revolt of the Netherlands (see LOW COUNTRIES), which he proved unable to subdue, was an irritation that his English and French enemies did their worst to inflame.

Cooperation between imperial and Spanish Habsburgs in the 17th century failed to maintain the hegemony that the dynasty had enjoyed in the 16th. For the imperial line, religious troubles in Germany and in central Europe went on even when the domestic conflict between the insane emperor Rudolf II and his brothers was over (1612); and the Bohemian insurrection of 1618 gave rise to that chain of wars involving the Austrian Habsburgs that, because it was prolonged until 1648, is known conventionally as the Thirty Years' War. For the Spanish Habsburgs, their truce of 1609 with the Dutch ended in 1621, whereupon the renewed conflict in the Netherlands became merged with the struggles of their Austrian cousins. The Peace of Westphalia (1648) finally abolished Habsburg sovereignty over the northern Netherlands, severely restricted the emperor's authority over the other German princes, and transferred the Habsburg lands in Alsace to France; however, the ordinance of 1627, whereby the Bohemian crown had been converted into a hereditary one for the Habsburgs, was permitted to stand.

The Thirty Years' War

The challenge of 17th-century France

France from the late 1620s had made the most of the Thirty Years' War to distress the Habsburgs of both lines, and peace with the imperial line did not prevent France from continuing its war against the Spanish until 1659, when by the Peace of the Pyrenees it obtained Gravelines, most of Artois, and part of Hainaut, together with some places south of Luxembourg.

The next 30 years saw the end of the Habsburg dynasty's claim to European hegemony in any real sense. The aggressions of Louis XIV of France, from 1667 onward, took territory after territory from the Spanish Habsburgs—large parts of Flanders, the rest of Artois, and other areas in the Netherlands, as well as the whole Franche Comté and, in 1684, the stronghold of Luxembourg—and demonstrated at the same time that the imperial Habsburgs, preoccupied as they were with the Turkish assault from Hungary, could not effectively defend the German frontier west of the Rhine. After being saved from the crisis of the Turkish siege of Vienna in 1683, the imperial Habsburgs did indeed obtain one dynastically significant success—the conversion, in 1687, of the Hungarian crown into an hereditary one for themselves—but by this time it was plain to Europe that the most formidable dynasty was no longer the Habsburg but the Bourbon. In the War of the Grand Alliance (1689–97) the rising powers that 100 years earlier had been Habsburg Spain's principal enemies and feeble France's most fluent encouragers, the Dutch and English, led those supporting the Habsburgs against Louis XIV.

Apart from the Bourbon ascendancy, there was a further reason for other powers to watch with jealous solicitude over the fate of Spain. The physical debility of Charles II of Spain was such that no male heir could be expected

The Spanish succession

to be born to him, and the question of his succession was one of great concern to the European powers. Up to 1699 it was understood that his crowns would pass to the electoral prince of Bavaria, Joseph Ferdinand, son of his niece Maria Antonia, daughter of the emperor Leopold I; and this arrangement was generally acceptable because, by transferring the Spanish inheritance to the Bavarian House of Wittelsbach, it would not necessarily upset the balance of power between the imperial Habsburgs and Bourbon France. In 1699, however, when Joseph Ferdinand died, the moribund Charles II's next natural heirs were the descendants (1) of his half-sister, who had married Louis XIV of France, and (2) of his father's two sisters, of whom one had been Louis XIV's mother and the other the emperor Leopold I's. Critical tension developed: on the one hand neither the imperial Habsburgs nor their British and Dutch friends could consent to their Bourbon enemy's acquiring the whole Spanish inheritance; on the other neither Bourbon France nor its British and Dutch enemies wanted to see an imperial Habsburg reunite in one pair of hands most of what the emperor Charles V had had in 1519. Charles II in the meantime regarded any partition of his inheritance as a humiliation to Spain: dying in 1700, he named as his sole heir a Bourbon prince, Philip of Anjou, the second of Louis XIV's grandsons. The War of the Spanish Succession ensued.

THE HABSBURG SUCCESSION IN THE 18TH CENTURY

To allay British and Dutch misgivings, Leopold I and his elder son, the future emperor Joseph I, in 1703 renounced their own claims to Spain in favour of Joseph's brother Charles, so that he might found a second line of Spanish Habsburgs distinct from the imperial; but when Joseph I died, leaving only daughters, in 1711, and was succeeded by his brother as emperor (Charles VI) and as ruler of the Austrian, Bohemian, and Hungarian lands, the British and the Dutch lost interest in making him king of Spain and together began serious negotiations with France. Their Treaties of Utrecht (1713), which recognized the Bourbon accession to Spain and to Spanish America, virtually forced the hand of the reluctant Charles, who made peace with France by the Treaty of Rastatt in 1714: out of the whole inheritance of the Spanish Habsburgs, he had finally to content himself with the southern Netherlands and with the former Spanish possessions on the mainland of Italy, together with Mantua (annexed by him in 1708) and Sardinia. Sardinia, however, was exchanged by him in 1717 for Sicily, which the peacemakers of Utrecht had assigned to the House of Savoy. With characteristic obstinacy, Charles remained technically at war with Bourbon Spain until 1720, when an armistice was declared (formal recognition of the Bourbon accession came only in 1725).

Meanwhile the extinction of the Spanish Habsburgs' male line and the death of his brother Joseph left Charles, in 1711, as the last male Habsburg. He had therefore to consider what should happen after his death. No woman could rule the Holy Roman Empire, and furthermore the Habsburg succession in some of the hereditary lands was assured only to the male line. In order, therefore, to secure the indivisibility of his Habsburg inheritance he issued his famous Pragmatic Sanction of April 19, 1713, prescribing that, in the event of his dying sonless, the whole inheritance should pass (1) to a daughter of his, according to the rule of primogeniture, and thence to her descendants; next (2) if he himself left no daughter, to his late brother's daughters, under the same conditions; and finally (3) if his nieces' line was extinct, to the heirs of his paternal aunts. The attempt to win general recognition for his Pragmatic Sanction was Charles VI's main concern from 1716 onward (his baby son died in that year). By 1738, at the end of the War of the Polish Succession (in which he lost both Naples and Sicily to a Spanish Bourbon but got Parma and Piacenza for the Habsburgs in compensation), he seemed to have won his point: Saxony, Bavaria (grudgingly and with an express reservation), Spain, Russia, Prussia, Hanover-England, and finally France (with a reservation about third-party rights) had all, in one way or another, acknowledged the Pragmatic Sanction. His hopes were illusory: less than two months after his death, in 1740, his daughter Maria Theresa had to face a Prussian invasion of Silesia, which unleashed the War of the Austrian Succession. Bavaria then promptly challenged the Habsburg position in Germany; and France's support of Bavaria encouraged Saxony to follow suit and Spain to try to oust the Habsburgs from Lombardy. Great Britain came, late enough, to support Maria Theresa rather out of hostility toward France than out of loyalty to the Pragmatic Sanction.

The Pragmatic Sanction and the Austrian succession

The challenge of Prussia

HABSBURG–LORRAINE

The War of the Austrian Succession cost Maria Theresa most of Silesia, part of Lombardy, and the duchies of Parma and Piacenza (Treaty of Aix-la-Chapelle, 1748) but left her in possession of the rest of her father's hereditary lands. Moreover, her husband, Francis Stephen of Lorraine, who in 1737 had become hereditary grand duke of Tuscany, was finally recognized as Holy Roman emperor, with the title of Francis I. He and his descendants, of the House of Habsburg–Lorraine (see Table 3), are the dynastic continuators of the original Habsburgs.

The peace of 1748 did not last long. Prussia was not satiated by the seizure of Silesia from the Habsburgs, and they in turn were even more determined to recover Silesia than anxious to ensure the protection of their outlying possessions in the Netherlands against the continuing danger of French attack. The so-called Diplomatic Revolution, which preceded the Seven Years' War of 1756–63, was the product, basically, of these situations: finding that their former British friends were more interested in conciliating Prussia than in abetting Austro-Russian plans for destroying it, the Habsburgs played their part in the "reversal of alliances" by achieving—without territorial profit—a reconciliation with France, hitherto their longest-standing enemy. An Austro-French entente was subsequently maintained until 1792: the marriage of the archduchess Marie-Antoinette to the future Louis XVI of France (1770) was intended to confirm it.

To secure their imperial status in Germany against Prussian enterprises, the Habsburgs exerted themselves to consolidate and to expand their central European bloc of territory. For this purpose Tuscany and the Netherlands were practically irrelevant. Tuscany in fact was kept separate from the ancient Habsburg inheritance: when the emperor Francis I died (1765), his eldest son, the emperor Joseph II, became coregent with his mother of the Austrian dominions, but Joseph's brother Leopold became grand duke of Tuscany; and similarly when Leopold succeeded to Joseph's titles (1790), his own second son succeeded to Tuscany as Ferdinand III. Thereafter the Tuscan branch of the Habsburgs remained distinct from the senior or imperial line.

The northeastward expansion of Habsburg central Europe, which came about in Joseph II's time, was a result not so much of Joseph's initiative as of external events: the First Partition of Poland (1772), which gave him Galicia and Lodomeria, was a Russo-Prussian arrangement disgusting to his conscientious mother, who remembered Silesia; and his subsequent acquisition of Bukovina (1775), geopolitically logical though it was as bridging a gap between his Transylvanian and his new Galician lands, was a side effect of the Russo-Turkish Treaty of Küçük Kaynarca (1774).

Joseph II was considerably more interested in westward expansion, over Bavaria, which would have both strengthened his western frontier strategically and enhanced his status among the German princes politically. Prussia's forceful opposition, however, reduced his gains in the War of the Bavarian Succession to the Innviertel (1779) and frustrated his plan for ceding the Netherlands to the House of Wittelsbach in exchange for Bavaria five years later (1784).

The French Revolutionary and Napoleonic Wars brought a kaleidoscopic series of changes. Three were clearly significant for the future of the House of Habsburg: (1) the formal dissolution of the Holy Roman Empire in 1806, in anticipation of which Leopold II's successor Francis II had in 1804 begun to style himself "hereditary emperor of Austria," a title that, as Francis I, he could retain come

The Austrian Empire

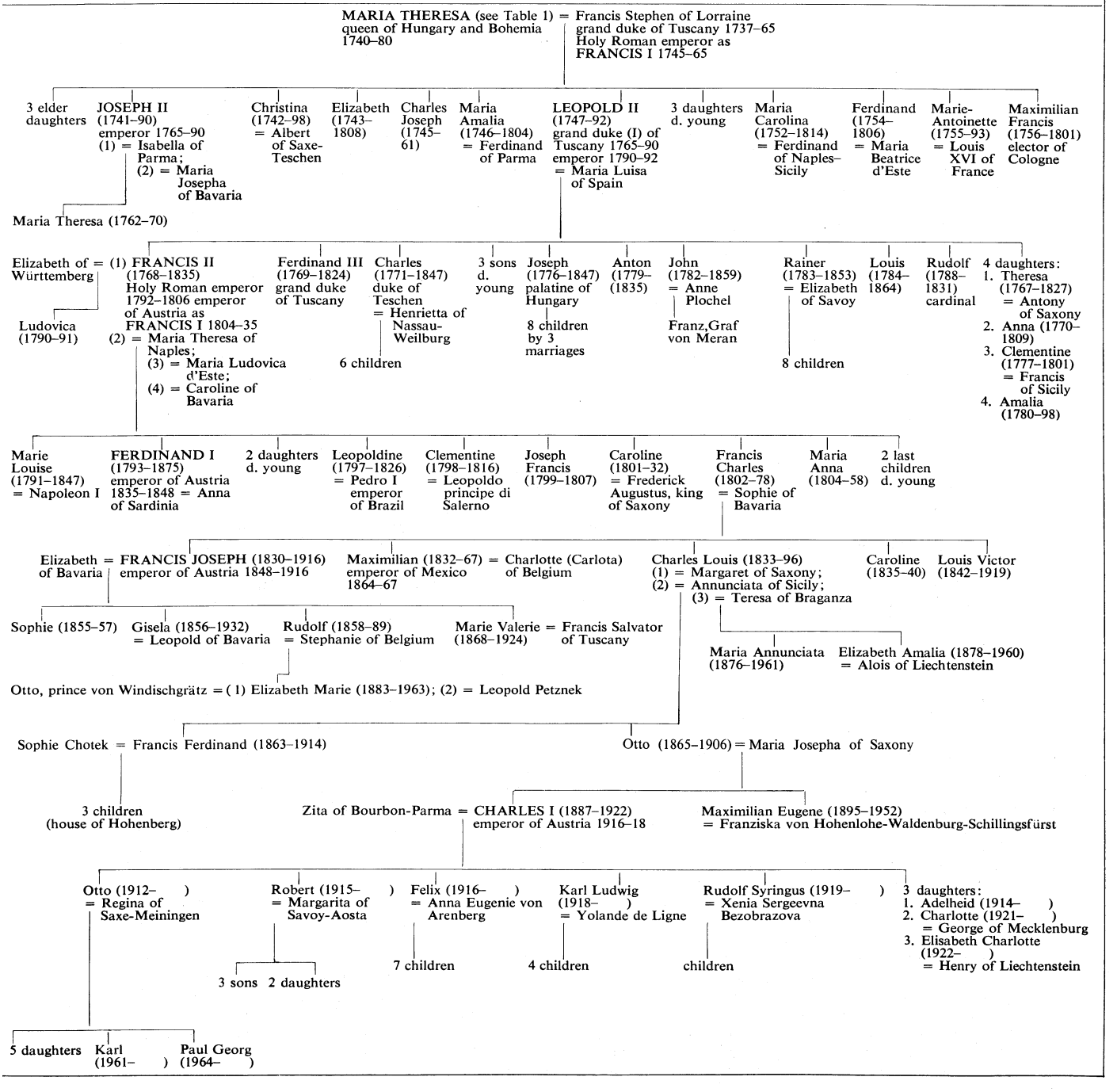
what might; (2) the definitive renunciation of the southern Netherlands by the Habsburgs in 1797; and (3) the awakening of the spirit of nationalism in the modern sense.

On Napoleon's downfall the Congress of Vienna (1814–15) inaugurated the Restoration, from which the battered House of Habsburg naturally benefitted. Francis I of Austria recovered Lombardy (lost in 1797), Venetia and Dalmatia (both of them acquired in 1797 but lost in 1809), and Tirol (lost also in 1809); Ferdinand III of Tuscany recovered his grand duchy; another Habsburg was recognized as sovereign duke of Modena, because his father, a brother of the Holy Roman emperors Joseph II and Leopold II, had in 1771 married the heiress of the House of Este; and Napoleon's Habsburg consort, Marie Louise, received the duchies of Parma and Piacenza for her lifetime (after which they were to revert to the Bourbons). The territory of Salzburg, which the Habsburgs had acquired in 1803 but lost to Bavaria in 1809, was finally restored to Austria in 1816. Though the Congress of Vienna did not restore Austrian rule over "Western Galicia" (the Habsburgs' share under the Third Partition of Poland in 1795, lost likewise in 1809), a small part of that area, namely the territory of Cracow, was annexed by Austria in 1846.

Nationalism against the dynasty

The history of the House of Habsburg for the century following the Congress of Vienna is inseparable from that of the Austrian Empire, a bastion of monarchical conservatism that the forces of nationalism—German, Italian, Hungarian, Slav, and Romanian—gradually eroded. The first territorial losses came in 1859, when Austria had to cede Lombardy to Sardinia–Piedmont, nucleus of the emergent kingdom of Italy, and could do nothing to prevent the same power from dispossessing the Habsburgs of Tuscany and of Modena. Next, the Seven Weeks' War of 1866, in which Prussia, exploiting German nationalism, was in alliance with Italy, forced Austria both to renounce its hopes of reviving its ancient hegemony in Germany and to cede Venetia. After this disaster the Habsburg emperor

Table 3: The House of Habsburg–Lorraine in Austria



Francis Joseph took a step intended to consolidate his "multinational empire": in 1867, to conciliate Hungary, he granted to that kingdom equal status with the Austrian Empire in what was henceforth to be the Dual Monarchy of Austria–Hungary. The result, however, was that the Magyars, jealous of their unique parity with the Germans and of their superiority over the non-Magyar peoples of their kingdom, rejected any suggestion of conciliating the Slavs and the Romanians of the Dual Monarchy by similar measures. The ardent German nationalists of the Austrian Empire, as opposed to the Germans who were simply loyal to the Habsburgs, took the same attitude as did the Magyars.

Remote from Austria's national concerns but still wounding to the House of Habsburg was the fate of Francis Joseph's brother Maximilian: set up by the French as emperor of Mexico in 1864, he was executed by a Mexican firing squad in 1867. No less grievous to the dynasty and of more concern to Austria–Hungary was the suicide of the crown prince Rudolf in 1889, though his fitness for the imperial and royal succession was questionable; and the scandalous misconduct of certain archdukes and archduchesses, in the imperial and in the Tuscan lines alike, further impaired the Habsburgs' personal prestige. The assassination of Francis Joseph's Wittelsbach consort Elizabeth in 1898 was to be followed in less than two decades by an assassination of far greater consequence.

In 1878 Austro-Hungarian forces had "occupied" Bosnia–Hercegovina, which belonged to decadent Turkey. In 1908 that territory had been formally annexed to Austria–Hungary, in a manner that was outrageous not only to Serbia (which coveted Bosnia for itself) but also to Serbia's patron, Russia. Visiting the Bosnian capital, Sarajevo, in 1914, the archduke Francis Ferdinand, heir presumptive to the Dual Monarchy (and incidentally legatee, from 1875, of the rights of the House of Austria–Este to Modena), was shot to death by a nationalist Serb. A month later the First World War was beginning.

World War I and after

World War I led to the dismemberment of the Habsburg Empire. While Czechs, Slovaks, Poles, Romanians, Serbs, Croats, Slovenes, and Italians were all claiming their share of the spoil, nothing remained to Charles, the last emperor and king, but "German" Austria and Hungary proper. On November 11, 1918, he issued a proclamation recognizing Austria's right to determine the future form of the state and renouncing for himself any share in affairs of state, and on November 13 he issued a similar proclamation to Hungary. Even so, he did not abdicate his hereditary titles either for himself or for the Habsburg dynasty. Consequently the national assembly of the Austrian Republic passed the "Habsburg Law" of April 3, 1919, banishing all Habsburgs from Austrian territory unless they renounced all dynastic pretensions and loyally accepted the status of private citizens. In Hungary, however, the collapse of the republican regime at the end of 1919 raised strong royalist hopes of a Habsburg restoration, and after the conclusion of the Treaty of Trianon (June 1920) Charles twice tried to return (March and October 1921). Under pressure from the other European powers, especially those of the Little Entente (Czechoslovakia, Yugoslavia, and Romania), the Hungarian parliament on November 3, 1921, decreed the abrogation of Charles's sovereign rights and of the Pragmatic Sanction.

Habsburg property rights in Austria, forfeited under the law of 1919, were restored in 1935 but withdrawn again by the German chancellor Adolf Hitler in 1938. After World War II the Allied Control Council in Austria in January 1946 declared that it would support the Austrian government in measures to prevent any return of the Habsburgs, and the law of 1919 was written into the Austrian State Treaty of 1955. In June 1961 the Austrian government rejected an application by the archduke Otto, head of the House of Habsburg, to be allowed to return to Austria as a private citizen, but in 1963 the administrative court of Austria ruled that Otto's application was legal. Because of Socialist opposition to his return, however, he was not granted a visa until June 1966 after the People's Party had won a majority in that year's general election. (J.R.-S.)

BIBLIOGRAPHY. For the English reader, the most comprehensive introduction to the subject remains ARCHDEACON WILLIAM COXE, *History of the House of Austria,* 3rd ed., 4 vol. (1847–53, reprinted 1971), a neglected work covering Habsburg history from 1218 to 1848. ADAM WANDRUSZKA, *The House of Habsburg: Six Hundred Years of a European Dynasty* (1964, reprinted 1975; originally published in German, 1956, reprinted 1982), covers 600 years more briefly but most authoritatively. Works in English on the Habsburgs in the 19th and 20th centuries vary in scope and in quality: CARLILE A. MACARTNEY, *The Habsburg Empire, 1790–1918* (1968, reprinted with corrections 1971), provides a masterly survey by an expert; ALAN J.P. TAYLOR, *The Habsburg Monarchy, 1809–1918: A History of the Austrian Empire and Austria-Hungary,* new ed. (1948, reprinted 1976), is rather severely critical of the dynasty's failings; EDWARD CRANKSHAW, *The Fall of the House of Habsburg* (1963, reprinted 1983), together with his lavishly illustrated *The Habsburgs: Portrait of a Dynasty* (1971), makes a brilliant and easily readable vindication of the last emperor. ARTHUR J. MAY, *The Hapsburg Monarchy, 1867–1914* (1960), is fair and scholarly, with a good bibliography, and is continued in his *The Passing of the Hapsburg Monarchy, 1914–1918,* 2 vol. (1966). ZBYNĚK A.B. ZEMAN, *The Breakup of the Habsburg Empire 1914–1918* (1961, reprinted 1977), gives a just account of a subject usually misrepresented. For genealogical detail, see WALTHER MERZ, *Die Habsburg* (1896), which contains 19 tables; and MICHEL DUGAST ROUILLE, *Les Maisons souveraines de l'Autriche, Babenberg, Habsbourg, (Habsbourg-d'Espagne), Habsbourg-Lorraine, (Lorraine)* (1967), also well tabulated and with illustrations. For the physical heredity, see OSWALD RUBBRECHT, *L'Origine du type familial de la maison de Habsbourg* (1910); WILHELM STROHMAYER, *Die Vererbung des Habsburger Familientypus* (1937); and JOHN LANGDON-DAVIES, *Carlos, the Bewitched* (1963; U.S. title, *Carlos, the King Who Would Not Die,* 1963). Other studies include DOROTHY G. MCGUIGAN, *The Habsburgs* (1966); VICTOR L. TAPIÉ, *The Rise and Fall of the Habsburg Monarchy* (1971; originally published in French, 1969); HUGH R. TREVOR-ROPER, *Princes and Artists: Patronage and Ideology at Four Habsburg Courts, 1517–1633* (1976), an illustrated study; SARAH GAINHAM, *The Habsburg Twilight: Tales from Vienna* (1979), a portrait of important figures of Vienna in the late 19th century; ROBERT A. KANN, *A History of the Habsburg Empire, 1526–1918* (1974, reprinted 1980); HELMUT G. KOENIGSBERGER, *The Habsburgs and Europe, 1516–1660* (1971), essays on the international impact of the Habsburgs; JOHN LYNCH, *Spain Under the Habsburgs,* 2 vol., 2nd ed. (1981).

Hamburg

The Free and Hanseatic City (Freie und Hansestadt) of Hamburg, on the Elbe River, is the second smallest of the 16 *Länder* (states) of Germany, with a territory of only 292 square miles (755 square kilometres). It is also the most populous city in Germany after Berlin and one of the largest and busiest ports in Europe. The official name, which covers both the *Land* and the town, reflects Hamburg's long tradition of particularism and self-government. Hamburg and Bremen (the smallest of the *Länder*) are, in fact, the only German city-states that still keep something of their medieval independence. The characteristic individuality of Hamburg has been proudly maintained by its people so that, in many spheres of public and private life, the city's culture has retained its uniqueness and has not succumbed to the general trend of standardization.

Hamburg, nonetheless, is a cosmopolitan city in its outlook. Although comparatively few foreigners live there, many pass through it. The city has dealings with a large number of nations, and it has more consulates than any other city in the world, except New York City. Shipping and trade have been Hamburg's lifeblood for centuries. Not surprisingly, its harbour has remained the city's most important feature.

Among Hamburg's many other facets are a network of canals reminiscent of Amsterdam; lakes, parks, and verdant suburbs full of gracious houses; elegant shopping arcades; richly endowed museums; and a vibrant cultural life. These are among the attractions that have contributed to a growing tourist industry. Although it was badly damaged during World War II, Hamburg has succeeded in maintaining a sense of old-world grace alongside its thriving commercial life.

This article is divided into the following sections:

Physical and human geography 472
- The landscape 472
  - Site
  - The city layout
  - Architecture
  - Climate
- The people 473
- The economy 473
  - Industry
  - Trade
  - Transportation
- Administration and social conditions 473
  - Government
  - Education
- Cultural life 473

History 474
- Early settlement and medieval growth 474
- Evolution of the modern city 474

Bibliography 475

Physical and human geography

THE LANDSCAPE

Site. Hamburg stands at the northern extremity of the Lower Elbe Valley, which at that point is between five and eight miles (eight and 13 kilometres) wide. To the southeast of the old city, the Elbe divides itself into two branches, the Norderelbe and the Süderelbe; but these branches meet again opposite Altona, just west of the old city, to form the Unterelbe, which flows into the North Sea some 65 miles downstream from Hamburg. Two other rivers flow into the Elbe at Hamburg—the Alster from the north and the Bille from the east.

The city layout. The nucleus of the city is the Altstadt (Old Town), the former medieval settlement, bounded by the harbour and by a string of roads that follow the line of the old fortifications. Within this core there are few great buildings to remind the visitor of the city's thousand-year history apart from the five principal churches—Sankt Jacobi, Sankt Petri, Sankt Katharinen, Sankt Nikolai, and Sankt Michaelis—and none of these is in its original condition. Fire has destroyed almost all the older residences and warehouses, and what was left untouched by conflagration has often been rebuilt for contemporary purposes. There are, however, a few scattered survivals of older buildings. Moreover, the layout of the old city centre can still be detected in some of the ancient street names and in the *Fleete* (canals), which connect the Alster with the docks on the Elbe. One of the best views of the inner city is to be enjoyed from the Lombardsbrücke (Lombard Bridge), whence the towers of the five churches can be seen rising high against a skyline that is still relatively harmonious despite the presence of modern skyscrapers.

At the heart of Hamburg is a lake, measuring 455 acres (184 hectares), formed by the damming of the Alster and divided by the Lombardsbrücke into the Binnenalster (Inner Alster) and the Aussenalster (Outer Alster). Around the latter are elegant suburbs such as Rotherbaum, Harvesterhude, and Uhlenhorst. Many waterways, navigable by pleasure boats, run into the Aussenalster.

Architecture. The last intact ensemble of traditional Hamburg architecture is to be found in the Deichstrasse, one side of which backs onto the Nikolai canal. Its tall, narrow houses, resembling those of Amsterdam, were originally built from the 17th through the 19th century. It was in one of them, number 42, now a restaurant, that the devastating fire of 1842 broke out. Afterward the houses were rebuilt in the old style. Today the street is a protected area, and in recent years it has undergone extensive restoration. Many traditional restaurants are found there.

Another survival of older architecture is in the Krameramtswohnungen, near Sankt Michaelis. Consisting of two half-timbered brick buildings on either side of a narrow courtyard, it was built as a series of dwellings for the widows of shopkeepers and is the only surviving 17th-century construction of its kind in the city. Thoroughly restored between 1971 and 1974, it now forms a delightful secluded alleyway housing a restaurant, small shops, and a branch of the Museum für Hamburgische Geschichte (Museum of Hamburg History).

Of Hamburg's five great churches, the most imposing is probably Sankt Michaelis, an 18th-century Baroque-style Protestant church with a rich white-and-gold interior. It was destroyed by fire in 1906, rebuilt, devastated again during World War II, and restored yet again after the war.

The prosperous years 1890–1910 brought an abundance of fine architecture, examples of which can be seen in the spacious and elegant patrician houses around the Aussenalster. Many of these are now occupied by consulates. Another period of architectural flowering came in the 1920s and 1930s when there was a revival of the use of the traditional north German dark red brick as a building material, led by the architects Fritz Höger and Fritz Schumacher. A good example is Höger's Chilehaus, a massive office building constructed between 1922 and 1924.

More recently Hamburg has acquired its quota of starkly functional modern buildings, such as the Congress Centrum (Congress Centre; opened 1973) and the Fernsehturm (Television Tower), 271.5 metres (891 feet) high, but there is now a strong tendency to renovate old houses

rather than to demolish and build afresh. Thus the townscape of Hamburg as a whole has a human quality lacking in many German cities.

Climate. Hamburg has mild winters, late springs, relatively cool summers, high humidity, and frequent fog. The mean winter temperature is 34.2° F (1.2° C), and the mean summer temperature is 62.4° F (16.9° C).

THE PEOPLE

In 1965 the city had about 1,850,000 inhabitants, but since then the population has been slowly decreasing. More than three-fourths of the residents are Protestants, and the remainder are predominantly Roman Catholic. There is a small Muslim community, which includes many Turkish *Gastarbeiter* ("guest workers"), and the Jews, of whom there had been 27,000 in 1933 (when Hitler took power), now number only about 1,000.

THE ECONOMY

Industry. Having absorbed Altona, Harburg, and Wandsbek in 1937, Hamburg has become Germany's major industrial city. All processing and manufacturing industries are represented there. Hamburg treats most of the country's copper supplies, and the Norddeutsche Affinerie, on Veddel, is Europe's second largest copperworks. The chemical, steel, and shipbuilding industries are also important, although shipbuilding has declined as a result of competition from Japan and Korea. Hamburg is also the most important centre in Germany, after Berlin, for newspaper and periodical publishing. Three nuclear plants, at Krümmel, Stade, and Brunsbüttel, provide power at a reasonable cost to the industries bordering the Unterelbe and to parts of Hamburg.

Industrial importance to Germany

Trade. In the period of German partition, Hamburg handled more than half of West Germany's foreign trade, not only in the form of shipping cargo but also as rail and airfreight. Chief among imports are vegetable oils and fats, tea, coffee, petroleum, tropical fruit, and uncured tobacco. Exports include machinery, electrotechnical products, processed petroleum fuel and lubricants, copper, and pharmaceutical products.

The greatest economic centre of Germany, Hamburg since 1960 has become the site of first-class trade fairs. Many of the fairs and conventions are held at the Ernst-Merck-Halle exhibition grounds, located south of the Planten un Blomen park. An especially popular event is the international boat show, held each winter.

Transportation. The harbour is Hamburg's "gateway to the world." More than 15,000 ships from over 100 countries pass through it each year. The city's Übersee-Zentrum is the world's largest roofed warehouse, and the Waltershof container terminal is the largest of its kind on the continent.

The port of Hamburg

Harbour and city are well served by the German railway network, and the city has a good system of buses and underground trains. To relieve the central city from long-distance traffic, a tunnel was built (opened in 1977) under the Elbe as a part of the Stockholm–Lisbon highway.

The airport of Hamburg–Fuhlsbüttel, which dates from 1911, is one of the oldest in Europe. It has two runways from which even the largest jet-propelled aircraft can still take off.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. According to the constitution of June 6, 1952, legislative authority is vested in the Bürgerschaft (State Parliament), which comprises 120 members elected for a four-year term.

The Bürgerschaft elects the government, the Senat, which is organized on a collegiate basis; the president, or *erster Bürgermeister* ("first mayor"), is elected by the Senat itself annually, although in practice each incumbent normally stays in office at least four years. The Senat as a whole represents the Free and Hanseatic City of Hamburg in its dealings with the other federal *Länder,* with the federal government, and with foreign states. Each senator is responsible for a particular department, but administrative problems of a local nature are delegated to the district offices and to the local authorities.

The magnificent Rathaus (City Hall), where the Senat and the Bürgerschaft meet, in the centre of the city near the Binnenalster, was built late in the 19th century in the Neo-Renaissance style.

Hamburg's coat of arms displays a three-towered castle, intended to represent the Hammaburg, silver (argent), on a red (gules) field, its design being derived from the city's great seal of 1241. The state flag likewise shows a white castle on a red field. The state's anthem, "Stadt Hamburg an der Elbe Auen" ("City of Hamburg by the Meadows of the Elbe"), was written by G.N. Bärmann in 1828 and set to music by Albert Methfessel.

© Mike Yamashita—Aspect Picture Library



The Rathaus on the Alsterfleet in Hamburg.

Education. The Universität Hamburg, founded in 1919, is one of the largest in Germany, with some 46,000 students and faculties covering virtually every discipline except certain technological subjects. A second university, the Technische Universität Hamburg-Harburg, began classes in 1982. Hamburg also has state schools for music and interpretative art and for the sculptural arts, as well as some 250 research centres covering such areas as hydrography, oceanography, tropical medicine, shipbuilding, economics, meteorology, and particle acceleration.

CULTURAL LIFE

Among Hamburg's six principal museums, the Kunsthalle, founded in 1868 by Alfred Lichtwark, an outstanding patron of artists, is one of Europe's most remarkable galleries. It is particularly notable for its collection of 19th- and 20th-century works, including many of the German Romantic school. The Museum für Kunst und Gewerbe (Museum of Art and Crafts), founded in 1877 by the jurist Justus Brinckmann, has one of the most significant

collections of ancient artifacts in Germany and is also famous for its examples of Asian art and of Jugendstil (Art Nouveau). The Museum für Hamburgische Geschichte, which has grown from a collection of local antiquities started in 1839, contains a wide range of exhibits, from costumes to parts of old buildings and from architect's drawings to models of ships, shown in such a way as to present an impressive conspectus of the state's history. The Museum für Völkerkunde (Museum of Ethnology and Prehistory), founded in 1878, has impressive collections in its own fields. The Altonaer Museum, opened in 1863, specializes in north German subjects, with special attention to Schleswig-Holstein, and houses Germany's largest collection of old ships' figureheads. The Helms-Museum, in the Harburg district, is a local museum for the part of Hamburg south of the Elbe but also houses antiquities representing the prehistory and early history of the whole territory. The Ernst-Barlach-Haus, in Jenisch Park, was founded in 1961–62 by another great patron of the arts, Hermann F. Reemtsma, to make his private collection accessible to the public. Hamburg's once famous Zoological Museum was destroyed by bombs in 1943 after a century of existence.

Music and drama in Hamburg

The Hamburg Staatsoper, which dates from 1678, has won world renown. Its performances of classical and contemporary works bear comparison with those given by the great opera houses of Vienna, Milan, London, and New York City. The Deutsche Schauspielhaus, a leading theatre, enjoyed a particularly high reputation from 1955 to 1963, when Gustaf Gründgens directed and performed there. The Thalia-Theater, founded in 1843, with a multifaceted program that includes plenty of light entertainment, is popular with local audiences. All three establishments are generously funded by the state. The numerous other theatres include the tiny Piccolo-Theater and the Hansa-Theater, said to be the last genuine variety theatre in the German-speaking world. Plays of a local character or in Plattdeutsch (Low German) are performed in the Ohnsorg-Theater and sometimes in the Sankt Pauli-Theater, which dates from 1841 and is Hamburg's oldest playhouse.

The birthplace of Mendelssohn and Brahms, Hamburg has a sustained tradition of musical activity. Three great orchestras—the Philharmonische Staatsorchester, the Symphonie-Orchester des Norddeutschen Rundfunks, and the Hamburger Symfoniker—familiarize the public with classical and contemporary compositions. There are also groups specializing in chamber music, in choral performances, and in church music; and orchestras, choirs, singers, and instrumentalists from other parts of Germany or from abroad are also invited to Hamburg. The focal point of Hamburg's musical life is the Neo-Baroque Musikhalle, built in 1904–08 with money donated by the shipowner Carl Laeisz.

An important aspect of Hamburg's history is its prominence as a centre of newspaper and periodical publishing, which dates to the 17th century. By 1830 Hamburg had more newspapers than any other city in Germany. In the 1920s Berlin began its ascendancy as a press centre, and Hamburg fell into second place, which it still occupies. Its daily newspapers include the *Hamburger Abendblatt*, the *Hamburger Morgenpost*, and the *Bild-Zeitung*. In addition, a wide range of weekly and monthly magazines issues from the various publishing firms located in the city.

Sports

Sports have long been popular in Hamburg. The Hamburger Turnerschaft (1816) is Germany's oldest athletic club. The first German rowing club, founded in Hamburg in 1836, took part in 1837 in Germany's first official rowing race, against the English Rowing Club formed by members of Hamburg's then numerous English colony. The Hamburger Rennklub, for horse racing, was founded in 1852, and the North German Derby, first run in 1869, became an annual event, as the German Derby, from 1889 onward. Hamburg's first public football matches were played in 1881–82, after disputes about the rules of the game with the local Anglo-American Football Club. Another noteworthy event is the international tennis championship, which takes place every year in May.

An aspect of the city that cannot be ignored is its world-famous red-light district, centred on the Reeperbahn, where prostitution is legally permitted and kept under control by the police. Another famous Hamburg institution is the colourful Sunday-morning market held at the Fischmarkt by the harbour, where items ranging from fish to secondhand household goods are available.

History

EARLY SETTLEMENT AND MEDIEVAL GROWTH

Hamburg's history begins with the Hammaburg, a moated castle of modest size, built in about AD 825 on a sandy promontory between the Alster and Elbe rivers. In 834, during the reign of the emperor Louis the Pious, the castle's baptistery became the seat of an archbishopric, and Archbishop Ansgar made the young city of Hamburg the base of his missions to the heathens of northern Europe. Vikings burned the city in 845, and the rebuilt Hamburg was burned down again eight times in the following 300 years. By the end of the 11th century, Hamburg's role as the spiritual metropolis of the north was over, and henceforth commerce rather than religion was to be the principal *raison d'être* of the city. Between 1120 and 1140 some trading businesses were installed, and the foundation of Lübeck, on the Baltic, by Adolf II, count of Holstein, further promoted the economic development of Hamburg as Lübeck's port on the North Sea. In the autumn of 1188 a group of Hamburg entrepreneurs received from their feudal overlord, Adolph III of Schauenburg (Schaumburg), count of Holstein, a charter for the building of a new town, adjacent to the old one, with a harbour on the Alster River and with facilities for the use of the Elbe River as an outer roadstead. On May 7, 1189, the emperor Frederick I Barbarossa confirmed Count Adolph's dispositions in a charter granting special trading rights, toll exemptions, and navigation privileges.

In the 13th century Hamburg grew steadily in both area and economic importance, owing to the development of the Hanse (an association of merchants trading in a particular area) into a widespread association of north German merchant cities, the great Hanseatic League, in which Hamburg's role was second only to Lübeck's. A major entrepôt for the trade between Russia and Flanders, Hamburg proceeded to safeguard the trade routes by acquiring tracts of land along the branches of the Elbe in the immediate vicinity of the town and also on the estuary farther downstream (Ritzebüttel, nucleus of the later Cuxhaven, was acquired by Hamburg in 1393). It thus came to control the use of the river and to be recognized as the protector, in the emperor's name, of navigation on its lower course. Some political complications arose with the death, in 1459, of the last Schauenburg count of Holstein, since his princely rights in Germany passed thereafter to the royal house of Denmark, but Hamburg scarcely recognized Danish suzerainty in any but a formal way.

Role in the Hanseatic League

EVOLUTION OF THE MODERN CITY

Toward the end of the Middle Ages, the Hanseatic League gradually dissolved. Hamburg then went its own way and by 1550 had surpassed even Lübeck in economic importance. A stock exchange was founded in 1558 and the Bank of Hamburg in 1619; a convoy system for shipping was inaugurated in 1662, Hamburg's merchantmen being the first to be escorted on the high seas by men-of-war. About the same time marine insurance was first introduced into Germany. There were two causes for this new ascendancy: first, the wars of religion in the Low Countries in the second half of the 16th century had prompted many Dutch merchants to emigrate to the Unterelbe (Lower Elbe) region, with the result that Hamburg was henceforth to be the focus of their already established international commerce; second, the city had been so efficiently fortified in the decade 1616–25 that it could pursue its business untroubled throughout the worst crises of the Thirty Years' War (1618–48). By the end of the 17th century, Hamburg, with 70,000 inhabitants, was the largest city in Germany after Cologne.

The Treaty of Gottorp, concluded with the Danes on May 27, 1768, released Hamburg from theoretical subjec-

tion to the king of Denmark and so paved its way to being acknowledged, in 1770, as an "immediate" imperial city of Germany (that is, having no overlord other than the emperor). In addition, the treaty ceded to Hamburg the islands, from Veddel to Finkenwerder, that lay between the city and the left banks of the Elbe River and that, a century later, were to be the site of new docks. Hamburg, however, was not to enjoy its new advantage for long: the Napoleonic Wars overthrew the old order in Germany, and in 1810 the little state was annexed to Napoleon's French Empire.

After Napoleon's downfall (1814–15), Hamburg became a member state of the German Confederation, with the designation "Free and Hanseatic City of Hamburg" from 1819. Prosperity was quickly recovered, as Hamburg's trade was extended to newly opened territories in Africa, Asia, and the Americas. Even the great fire of May 1842, which devastated one-fourth of the city, did not check the booming economy, and the harbour was converted into one accessible at any time, without ships' having to depend on the state of the tides in the Elbe Estuary. Under the German Empire, founded in 1871, the political status of Hamburg was maintained, and development proceeded unchecked. The splendid Baroque houses of the densely populated Brook Island were demolished in the 1880s to make room for the warehouses of the new free port. By the end of the 19th century, in the course of which the population grew from 130,000 to 700,000, Hamburg had expanded far beyond its previous limits, absorbing such former suburbs as Sankt Pauli and Sankt Georg and spreading its tentacles into the countryside, toward Eimsbüttel, Eppendorf, Harvestehude, and Barmbek.

Hamburg entered the 20th century determined to maintain and to strengthen its position as "Germany's gateway to the world"; new docks and wharves were constructed on the left bank of the Elbe River. The outbreak of World War I in 1914 brought progress to a standstill, however. Hamburg's international trade collapsed, and its merchant fleet of 1,466 ships was virtually confined to port. After the war the victorious Allies demanded nearly all of Hamburg's ships by way of reparation from Germany.

The city between World Wars I and II

For many years after the war, Hamburg could undertake no further development because it had already exhausted all the potentialities of its territory. The Greater Hamburg Ordinance of January 26, 1937, changed this situation by allowing Hamburg to incorporate the neighbouring cities of Altona, Wandsbek, and Harburg, which until then had belonged to Prussia. The immediate prospect of expansion, with the development of these areas on a basis of large-scale planning, was shattered by the outbreak, in 1939, of World War II, during which repeated air raids demolished 55 percent of Hamburg's residential area and 60 percent of the harbour installations and killed 55,000 people. When the war ended in 1945, only the most strenuous efforts could supply the elementary needs for Hamburg's survival.

Reconstruction proceeded rapidly, however. Symptomatic of the city's postwar commercial efflorescence is the vast new business district City-Nord, built in the 1960s.

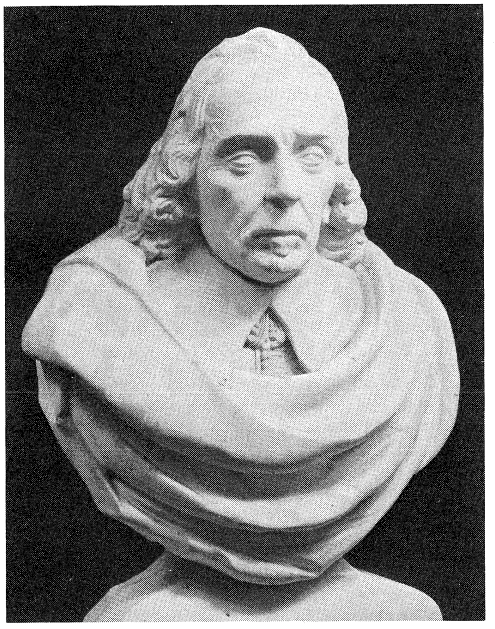
Hamburg's cherished traditions, together with its thriving business and cultural life and the energy of its inhabitants, make it one of the most vibrant cities in the world.

BIBLIOGRAPHY. BERNHARD MEYER-MARWITZ, *Das Hamburg Buch* (1981), a concise and lively book for the general reader; DAVID RODNICK, *A Portrait of Two German Cities: Lübeck and Hamburg* (1980), the city's history and social conditions; ECKART KLESSMAN, *Geschichte der Stadt Hamburg* (1981), a comprehensive and detailed study; HEINRICH REINCKE, *Hamburg: Ein Abriss der Stadtgeschichte von den Anfängen bis zur Gegenwart* (1926), a historical presentation; WILSON KING, *Chronicles of Three Cities—Hamburg, Bremen, Lübeck* (1914), an informative illustrated source; MARTIN CAIDIN, *The Night Hamburg Died* (1960), dealing with the events of World War II; and VOLKER PLAGEMANN (ed.), *Industriekultur in Hamburg* (1984), a well-illustrated symposium focussing on social and cultural aspects of the city's industrial history. The architecture of the city is explored in *Die Bau- und Kunstdenkmale der Freien und Hansestadt Hamburg,* ed. by GÜNTHER GRUNDMAN, 3 vol. (1953–68). (He.Th./C.A.McI.)

Harvey

A leading English physician of the first half of the 17th century, William Harvey achieved fame by his conclusive demonstration of the true nature of the circulation of the blood and the function of the heart as a pump. Functional knowledge of the heart and the circulation had remained almost at a standstill ever since the time of the Greco-Roman physician Galen—1,400 years earlier. Harvey's courage, penetrating intelligence, and precise methods were to set the pattern for research in biology and other sciences for succeeding generations, so that he shares with William Gilbert, investigator of the magnet, the credit for initiating accurate experimental research throughout the world.

By courtesy of Hempstead Parish Church, Essex; photograph, the Royal Academy of Arts, London



Harvey, marble bust by Edward Marshall. In Harvey Chapel, St. Andrews Church, Hempstead, Essex.

William Harvey was born on April 1, 1578, at Folkestone, Kent. His father, Thomas Harvey, was a prosperous businessman and a leading citizen of the small town. William, the eldest of nine children, was the only one to achieve special distinction in his career, but all his brothers were successful in business or at the royal court in London and among them amassed considerable wealth.

Career as physician and scientific innovator. Little is known of William Harvey's boyhood in the Kentish countryside. During the years 1588 to 1593 he was at the King's School attached to the cathedral at Canterbury. In his 16th year Harvey entered Gonville and Caius College, Cambridge, where he was awarded a scholarship in 1593. Although Harvey attended Caius College because of its special interest in educating doctors, his training was grossly inadequate. He was absent from the university for the greater part of his last year (1598–99) because of illness—probably malaria—but had received the B.A. degree in 1597. Determined to continue with medical training, he began a two-and-a-half-year course of study at the University of Padua, reputed to have the best medical school in Europe. His teacher was a celebrated anatomist, Hieronymus Fabricius ab Aquapendente, and it was in the now-famous oval Anatomy Theatre, still to be seen at the university, that Harvey first recognized the problems posed by the function of the beating heart and the properties of the blood passing through it.

Early theories of the heart and circulation

From the time of Aristotle in the 4th century BC it had been widely believed that the blood vessels contained both blood and air. Galen, the Greco-Roman physician, in the 2nd century AD proved that the arteries contained only blood but still believed that air entered the right side of the heart from the lungs. There was a general belief that the movement of the blood was by ebb and flow, an analogy being found in the movement of the sea. Galen's views on this are difficult to assess with exactitude, but it is apparent that he, like everyone else, had no conception of a circular movement of the blood, leaving the heart by one set of vessels, the arteries, and returning to it by another set, the veins. The main propulsive force initiating this oscillatory movement was supposed to be derived from a contracting of the arterial system, rather than by a pumping action of the heart. The blood in the veins was believed to be formed in the liver, passing to the right auricle (*i.e.*, one of the two upper chambers of the heart), and from there to the right ventricle (one of the two lower chambers), to make its way through holes in the septum, or partition, to the left side, where it met with blood from the arteries, which was mixed with air derived from the lungs. This was the extent of man's knowledge about the movement of the blood until 14 centuries later. Early in the 16th century the idea of a pulmonary circulation—that is, a circular motion of blood through heart and lungs—began to occur to some anatomists. In addition, the presence of a perforated septum was beginning to be questioned. In the middle of the 16th century a great anatomist, Andreas Vesalius, also working at Padua, first established accurate knowledge of human anatomy but was less interested in function. Several other medical investigators refined the anatomical knowledge of the heart. Realdus Columbus of Cremona, working as assistant to Vesalius, developed the idea of a pulmonary circulation, and this was made more definite by his pupil, Andreas Caesalpinus, though they still thought that the blood was distributed to the body by the great veins and their branches. Fabricius had a special interest in the anatomy of the veins and first described the system of valves found in them, but he was quite ignorant of their true function. In brief, there existed no convincing explanation of how the heart worked, and Harvey's logical mind remained unsatisfied.

His 28 months at Padua are only meagrely documented, but it is clear that he was outstanding among the students of his year. After receiving his diploma as doctor of medicine of Padua in April 1602, he returned to England. By the standards of the time he was fully trained in anatomy, the simpler functions of the human body, and in therapeutics based on the writings of Aristotle. He had had some clinical experience in the hospitals of Padua and Venice and was entitled to obtain a fellowship of the College of Physicians in London after passing through the preliminary stage of candidate for the higher qualification. At his first oral examination, in May 1603, he was given limited permission to practice medicine, but only after further examinations in April and August 1604 was he fully licensed to practice within the jurisdiction of the college—that is, in the London area.

Shortly after his return to England, Harvey married Elizabeth Browne, daughter of Lancelot Browne, physician to King James I and his queen, and a senior fellow of the college. The couple set up house in the parish of St. Martin's by Ludgate, not far from the College of Physicians; and, backed by Browne, Harvey then tried to obtain the appointment of physician to the Tower of London, where a number of distinguished men were imprisoned. Though he failed in this attempt, in 1607 he finally obtained a fellowship of the College of Physicians, which entitled him to seek an appointment as physician to one of the two great hospitals then serving London—St. Bartholomew's and

St. Thomas's. It may have been through his brother John, who had obtained employment in the King's household, that early in 1609 the King gave Harvey a recommendation for an appointment at St. Bartholomew's, which was conveniently near his house in St. Martin's. He was given the post of assistant physician, and, when the physician died in the summer of that year, Harvey succeeded him. The hospital at that time had about 200 beds for patients in 12 wards. Harvey's duties consisted of attending in the hall of the hospital to see the patients and prescribe for their treatment; he worked at least one day a week throughout the year and at any other time when specially needed. The physician was usually expected to live within the hospital precincts, but the rule was waived for Harvey since he lived not far away. He received an annual salary of £25 with £2 extra for his livery and a further £8 since he did not use the official residence. His colleagues were three surgeons and an apothecary in charge of the dispensary.

Appointment to St. Bartholomew's

Harvey held this office for 34 years, until 1643 when he was displaced for political reasons by Oliver Cromwell's party, then in power in London. These years saw the development and culmination of his active career as physician and scientific innovator. He developed a large private practice, attending many of the most distinguished citizens, including Sir Francis Bacon—and, about 1618, was made physician extraordinary to King James I, thus becoming a colleague of Sir Theodore Turquet de Mayerne, the senior court doctor. There can be no doubt that Harvey was for many years one of the most widely trusted doctors in England, although his unorthodox views on the circulation of the blood did injure his practice after their publication in 1628. Invariably courteous and regarded with affection and respect by his colleagues, he conducted his practice with common sense and honesty. Though advanced in his ideas of anatomy and physiology and scientific in his methods of research, he was inevitably conservative in the use of remedies. Very few potent drugs were known in his time, and accurate diagnosis was, more often than not, impossible, so that he never escaped from the influence of Aristotle, in whose principles he had been trained. He was the great protagonist of experimental biology but did not apply himself to this form of originality in therapeutics.

At the time of the King's last illness in 1625, de Mayerne was out of the country, and Harvey led the team of doctors in attendance. After the King's death it was rumoured that his favourite, the Duke of Buckingham, had contributed to the fatal outcome by applying remedies not approved by the doctors. He was actually accused of having poisoned the King, and an inquiry was ordered by Parliament in 1626. Harvey was the most important witness of several who contributed to exonerating the Duke from any direct responsibility. Charles I, the new king, continued Harvey's appointment as his personal physician and gave him a special award for the care he had given the last King. Charles's health remained good until the day of his execution, so that he rarely had need to consult his doctor. Nevertheless, Harvey became his close friend and was always in attendance on his journeys, such as his state visits to Scotland in 1633 and 1638. The King helped Harvey's scientific research by putting the deer in the royal parks at his disposal, and he delighted in showing the King anything of curiosity or scientific interest. At the same time, Harvey took his full share in the affairs of the College of Physicians, being constantly present at the meetings of the fellows and occupying all the official positions in the college hierarchy except that of president. His duties at court would not have allowed him to fill this position during his active years, and when it was offered to him in 1654 he was too old and ill to be able to accept. Yet it is clear from the college records that Harvey was always the man to whom his colleagues turned for advice. The physicians at this time had precedence over the other branches of the profession, and Harvey had a prominent part in maintaining this ascendancy over the surgeons, obstetricians, and apothecaries whenever they became restive under the authority of the college.

Scientific research

In spite of Harvey's activity in medical practice and college affairs, he spent much time in scientific research from the time of his return to England in 1604 until the beginning of the Civil War in 1642. His interest lay primarily in elucidating the facts of the movement of the heart and its relation to the circulation of the blood. Fabricius at Padua had opened his eyes to the value of comparative anatomy, and he was tireless in dissecting every kind of living thing, from insects, earthworms, reptiles, birds, and mammals up to man himself. He seized every opportunity to increase his knowledge of pathology through postmortem examinations and was an acute clinical observer of his patients, not omitting their psychology. Most of his scientific papers were destroyed by parliamentary soldiers during the Civil War, so that there is now no direct evidence of his methods. On the other hand, his lecture notes used from 1616 onward survive. In 1615 he was appointed to a college lectureship intended to cover all parts of medical knowledge, though each lecturer modified the course to suit his own interests. Harvey's manuscript, now in the British Museum, was entitled "Lectures on the Whole of Anatomy." It is written in a very bad hand in mixed Latin and English, and it is incomplete, lacking any account of the skeleton, the sense organs, and other systems. The systematic anatomy is enlivened by many references to comparative anatomy, morbid anatomy, and clinical observations, even naming individuals whom he had treated. It is evident that he wrote these notes before he had come to any conclusions about the circulation of the blood, so that they contain nothing that seriously questioned the authority of Galen. The only reference to his novel views is on a leaf inserted some years later, probably after 1628. Harvey held this lectureship until 1656.

Discovery of circulation of the blood. It is evident from his writings that Harvey reverenced Aristotle, even though he had to dismiss some of his teachings as absurd. He also valued the views of Galen, his predecessor in experimental physiology, and enlisted his support whenever he could do so. Yet Harvey depended essentially on reasoning from his own observations and experiments for proof of his contentions. During the 12 years after 1616, Harvey may have introduced some novelties into his lectures; but, by his own assertions, he demonstrated the results of his researches to his friends privately at the college. In 1628 he finally published his book, *Exercitatio Anatomica de Motu Cordis et Sanguinis in Animalibus* (*An Anatomical Exercise Concerning the Motion of the Heart and Blood in Animals*), a slender volume that established the true nature of the circulation of the blood. With Galen's help, he first disposed finally of the idea that the blood vessels contained air. He then elucidated the function of the valves in the heart in maintaining the flow of blood in one direction only when the ventricles contracted—on the right side to the lungs and on the left to the limbs and viscera. He proved that no blood passed through the septum, separating the two ventricles, and he explained the purpose of the valves in the larger veins in directing the return flow toward the heart. He showed that blood was expelled from the ventricles during contraction, or systole, and flowed into them from the auricles during expansion, or diastole. He proved that the arterial pulse was due to passive filling of the arteries by the systole of the heart and not by active contraction of their walls. He explained the purpose of the pulmonary circulation from the right ventricle through the lungs and back to the left auricle and ventricle. His only failure was in not demonstrating the connection of the arterial and venous systems in the tissues of the limbs by means of the smallest, or capillary, vessels. These he was unable to see, having no adequate form of microscope at his disposal. He was the first scientist to employ measurement of the content of the chambers of the heart and estimation of the total amount of blood in the body—that is, quantification.

Fame and criticism

Harvey's book made him famous throughout Europe, though the overthrow of so many time-hallowed beliefs attracted virulent attacks and much abuse from those who did not wish to believe the plain evidence of their senses. He refused to indulge in controversy and made no reply until 1649, when he published a small book answering the criticisms of a French anatomist, Jean Riolan. In this he reiterated some of his former arguments and utterly demolished Riolan's objections.

In 1636 King Charles dispatched a diplomatic embassy to the Holy Roman emperor Ferdinand II at Regensburg, Germany, in an attempt to establish the claim of his nephew, Prince Charles Louis, as Elector Palatine. Harvey was chosen as doctor to the mission and spent ten adventurous months of travel by land and water through territories ravaged by the Thirty Years' War, extending his journey by visits to Vienna, Prague, Venice, Rome, and Naples. At Nuremberg Harvey had a historical encounter with Caspar Hofmann, professor of medicine at the University of Altdorf, whom he attempted, at a public demonstration, to convince of the truth of his doctrine of the circulation. Though he did not succeed, Harvey behaved with great dignity and good temper in the face of obstinate blindness to demonstration of the facts.

At the start of the Civil War in 1642, Harvey was with the King and was in charge of the two princes, Charles and James, in the early stages of the Battle of Edgehill. When the King established his headquarters soon afterward at Oxford, Harvey remained with him and was given the position of warden of Merton College in 1645. Here he resumed his work on the development of the chick in hens' eggs and first met John Aubrey, antiquary and gossip, who afterward left a revealing account of Harvey in his *Brief Lives.* When the defeated King fled from Oxford to surrender himself to the Scots, Harvey joined him for a time at Newcastle but was forced to leave the King when he was handed over to the parliamentary army and was not allowed to go to him when he was imprisoned in the Isle of Wight. Harvey had never been much interested in politics but felt a deep personal regard for the King and after his execution in 1649 was a broken and unhappy man.

Yet two years later he published his second great book. After the publication of *De Motu Cordis,* the main achievement of Harvey's life, he had continued active research into the difficult subject of reproduction in animals. This led in 1651 to the publication of *Exercitationes de Generatione Animalium* (*Anatomical Exercitations Concerning the Generation of Animals*) through the persuasions of his younger friend Sir George Ent, a fellow of the college. The book contains much of historical and scientific interest, but Harvey's thought was greatly influenced by Aristotle. The book is mainly concerned with the development of the chick in hens' eggs, and Harvey insisted throughout that in all living things the origin of the embryo is to be found in the egg. He investigated also the embryology of deer, rejecting Aristotle's notion that menstrual blood played any part in the formation of the fetus; he also questioned whether or not semen had any influence. Having no microscope, he could not see the spermatozoa, which were not demonstrated until 1686 by Leeuwenhoek working in Holland with stronger lenses. Harvey remained uncertain of how fecundation of the ovum was accomplished and even suggested that it was by a kind of infection resembling the origin of infectious diseases. Aristotle had originated the theory of gradual formation of the embryo, part by part, as opposed to the idea of preformation, meaning that all the parts arose in miniature at the same time. Harvey agreed with Aristotle and crystallized the belief in the term epigenesis, though to him its meaning was extremely simple compared with all that is implied by it at the present time. Aristotle believed in the principle of spontaneous generation of primitive organisms; it is probable that Harvey did not support this belief, but his statements are equivocal, and his position remains uncertain.

Harvey's brothers had been successful merchants, and their advice, coupled with his skill as a doctor and his naturally austere habits, enabled him to accumulate considerable wealth. But he had become old and ill. He had met with so much opposition and disbelief that his passionate desire to establish scientific truth was partially unsatisfied. In his last years under Cromwell's Protectorate, he was regarded as a political "delinquent" owing to his long association with King Charles and was forced to spend most of his time lodging in one or another of his brothers' houses outside London. Though he corresponded with many distinguished foreign doctors, he was reluctant to engage in any further scientific research, saw few patients, and took little part in the affairs of the College of Physicians. He showed his regard for the fellows by giving them a new college building in 1652 with a library containing his own collection of books and presumably any remaining manuscripts. This was in use for less than 14 years, being destroyed in the Great Fire of London in 1666, so that very few of his books have survived to the present day. He suffered severe pain from gout and kidney stones and described himself to a correspondent as "not only ripe in years, but also a little weary and entitled to an honourable discharge" from further scientific argument. His last illness was brief. He awoke one morning partially paralyzed and unable to speak, probably owing to a cerebral thrombosis. He died in his 80th year on June 3, 1657, probably in his brother Eliab Harvey's house at Roehampton. He was buried in the family vault at Hempstead, an Essex village 50 miles (80 kilometres) from London. In 1883 he was reburied in a marble sarcophagus in the Harvey Chapel there, near a marble bust by Edward Marshall. This is a lifelike image of Harvey, better, probably, than any of the existing portraits of him in old age.

Last years

BIBLIOGRAPHY. SIR GEOFFREY KEYNES, *The Life of William Harvey* (1966, reissued 1978), is a full and definitive biography based on examination of contemporary sources, documented and illustrated, with eight appendixes; his *A Bibliography of the Writings of Dr. William Harvey, 1578–1657,* 2nd ed. (1953), is an account of all Harvey's books and of where they may be found; and his *The Portraiture of William Harvey* (1949) is a catalog of pictures, genuine and spurious, with reproductions. JOHN G. CURTIS, *Harvey's Views on the Use of the Circulation of the Blood* (1915), is an early study of the position of Harvey's work in the history of the knowledge of human physiology. For texts, see GWENETH WHITTERIDGE (ed.), *The Anatomical Lectures: Prelectiones Anatomie Universalis, De Musculis* (1964), a reliable transcription of Harvey's lecture notes, both in Latin and English, with a full discussion and interpretation, and WHITTERIDGE (trans.), *An Anatomical Disputation Concerning the Movement of the Heart and Blood in Living Creatures* (1976, trans. from Latin); see also WHITTERIDGE, *William Harvey and the Circulation of the Blood* (1970), an important study of the growth of Harvey's ideas. ARTHUR W. MEYER, *An Analysis of the De Generatione Animalium of Harvey* (1936), is a discussion of Harvey's second major publication, a work on animal reproduction and development; *De Generatione Animalium* is also treated in ELIZABETH B. GASKING, *Investigations into Generation, 1651–1828* (1967). WALTER PAGEL, *William Harvey's Biological Ideas: Selected Aspects and Historical Background* (1967), a well-documented historical analysis of Harvey's ideas on physiology and embryology, is continued in his *New Light on William Harvey* (1976). A good survey of Harvey's works is KENNETH D. KEELE, *William Harvey: The Man, the Physician, and the Scientist* (1965). Later studies include JEROME J. BYLEBYL (ed.), *William Harvey and His Age: The Professional and Social Context of the Discovery of the Circulation* (1979); and ROBERT G. FRANK, *Harvey and the Oxford Physiologists: A Study of Scientific Ideas* (1980), an analysis based upon diaries, letters, notebooks, manuscripts, and published scientific works.

(G.L.K./Ed.)

Havana

Havana (Spanish: La Habana) is the capital, major port, and leading commercial centre of Cuba. It also constitutes one of Cuba's 14 provinces. Located on the island's north coast, Havana, with more than 2,000,000 people, is the largest city in the Caribbean region and one of the great treasuries of historic colonial preserves in the Western Hemisphere. Prior to 1959, when Fidel Castro came to power, it was a Mecca for tourists from the United States, who were drawn by the city's many attractions, which included climate and nightlife in addition to history. In the next 30 years, however, despite its continued importance as the island's major economic hub, Havana lost much of its lustre as Castro's socialist government redirected the country's resources primarily toward the improvement of conditions in rural Cuba. Havana's upkeep and improvement thus lagged until the 1980s, when a major rehabilitation project began.

This article is divided into the following sections:

Physical and human geography 479
- The character of the city 479
- The landscape 479
  - The city site
  - Climate
  - The city layout
- The people 480
- The economy 481
  - Industry
  - Commerce and finance
  - Transportation
- Administration and social conditions 481
  - Government
  - Services
  - Health
  - Education
- Cultural life 481

History 482
- Foundation and early growth 482
- Development as a major New World port 482
- Alliance with the United States 482
- The city under Castro 482

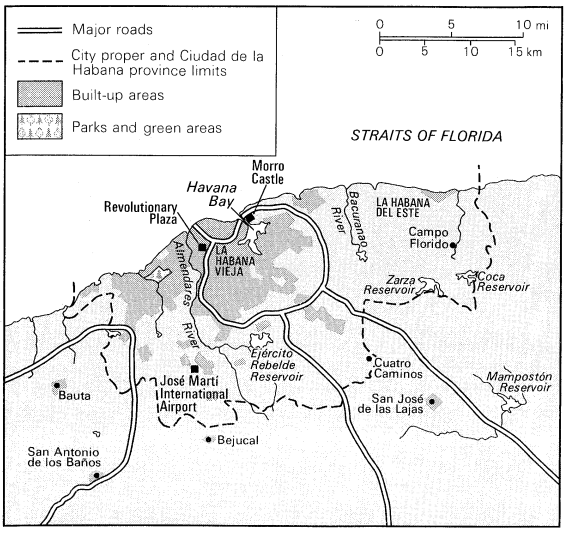
Bibliography 482

Physical and human geography

THE CHARACTER OF THE CITY

Havana's location along a magnificent deep-sea bay with a sheltered harbour made the city a prime location for economic development from Spanish colonial times in the early 16th century. Cuba is endowed with a number of such harbours, but Havana's on the north coast was prized above the others by the early Spanish colonizers. With land on both sides of the harbour, the port was easily defended. The early colonists erected a number of fortifications in the area that withstood most invaders. In colonial times Havana was the first landfall for Spanish fleets coming to the New World, and it became a staging area, first, for the conquest of the Americas by Spanish conquistadores and, later, for the economic and political domination of the hemisphere by Spain. The city early became a cosmopolitan centre with sprawling fortifications, cobblestone plazas, and buildings with ornamental facades and ornate iron balconies. Today's Havana mixes these structures with a variety of conventional modern buildings.

Havana's rich cultural milieu included not only Spaniards from diverse regions of the Iberian Peninsula but other European peoples as well. The small native Indian population of Cuba was not a significant factor in the Havana area and, in any case, was largely decimated in its early contact with the Spanish. The colonial years brought a large influx of black slaves from Africa who, after the end of slavery in the late 19th century, began flocking to Havana. Today's Havana is a mix of white Spanish stock, black ethnic groups, and significant mulatto strains.



Havana and surrounding area.

THE LANDSCAPE

The city site. The city extends mostly westward and southward from the bay, which is entered through a narrow inlet and which divides into three main harbours: Marimelena, Guasabacoa, and Atarés. The sluggish Almendares River traverses the city from south to north, entering the Straits of Florida a few miles west of the bay. The bay

The low hills on which the city lies rise gently from the deep blue waters of the straits. A noteworthy elevation is the 200-foot- (60-metre-) high limestone ridge that slopes up from the east and culminates in the heights of La Cabaña and El Morro, the sites of colonial fortifications overlooking the bay. Another notable rise is the hill to the west that is occupied by the University of Havana and the Prince's Castle.

Climate. Havana, like much of Cuba, enjoys a pleasant year-round climate that is tempered by the island's position in the belt of the trade winds and by the warm offshore currents. Average temperatures range from 72° F (22° C) in January and February to 82° F (28° C) in August. The temperature seldom drops below 50° F (10° C). Rainfall is heaviest in October and lightest from February through April, averaging 46 inches (1,167 millimetres) annually. Hurricanes occasionally strike the island, but they ordinarily hit the south coast, and damage in Havana is normally less than elsewhere in the country. Hurricanes

The city layout. Walls as well as forts were built to protect the old city, but by the 19th century Havana had already grown beyond the original barriers. The city first spread to the south and west. Expansion to the east was facilitated later by the construction of a tunnel under the entrance to the bay; such suburbs as La Habana del Este were subsequently able to be developed.

Several broad avenues and boulevards stretch across the city. One of the most picturesque is the Malecón, which extends southwestward along the coast from the port en- Major streets

trance to the Almendares River, under which it passes via a tunnel, emerging on the other side in Miramar as Avenida Quinta. Roughly paralleling the Malecón in the Vedado neighbourhood is Linea, another long avenue that passes under the river. Among other thoroughfares of note are Avenida del Puerto, Paseo Martí (or Prado), Avenida Menocal (Infanta), and Avenida Italia.

Contemporary Havana can essentially be described as three cities in one: Old Havana, Vedado, and the newer suburban districts. Old Havana, with its narrow streets and overhanging balconies, is the traditional centre of much of Havana's commerce, industry, and entertainment, as well as being a residential area. It is richly endowed with historic buildings, representing architectural styles from the 16th through the 19th century. Covering some three square miles and hugging the harbour, Old Havana includes Spanish colonial structures, towering Baroque churches, and buildings in Neoclassic style, as well as commercial property and less pretentious homes on the fringes.

Old Havana

To the north and west a newer section, centred on the uptown area known as Vedado, has become the rival of Old Havana for commercial activity and nightlife. This part of the city, built largely in the 20th century, contains attractive homes, tall apartments, and offices along wide, tree-lined boulevards and avenues. It is also the location of many hotels that before 1959 were frequented by U.S. tourists. Central Havana, sometimes described as part of Vedado, is mainly a shopping district that lies between Vedado and Old Havana.

A third Havana is that of the more affluent residential and industrial districts that spread out mostly to the west. Among these is Marianao, one of the newer parts of the city, dating mainly from the 1920s. Some of the suburban exclusivity was lost after the revolution, many of the suburban homes having been expropriated by the Castro government to serve as schools, hospitals, and government offices. Several private country clubs were converted to public recreational centres.

From colonial times Havana has been noted for its parks and plazas. Habaneros, as its residents are called, gather day and night under the sprawling trees of these many green areas. Through colonial times and almost to the end of the 19th century, the Plaza de Armas in Old Havana was the centre of Cuban life. Its most famous building, completed in 1793, is the Palace of the Captains General, an ornate structure that housed the Spanish colonial governors and, from 1902, three Cuban presidents. The building is now a museum.

The Plaza de Armas

In the 1980s many parts of Old Havana, including the Plaza de Armas, became part of a projected 35-year multimillion-dollar restoration project. The government sought to instill in Cubans an appreciation of their past and also to make Havana more enticing to tourists in accordance with the government's effort to boost tourism and thus increase foreign exchange.

One of the first buildings to be restored was the Cathedral of Havana, the church of Havana's patron saint, San Cristóbal (St. Christopher); it was constructed in the 18th century by the Jesuit order. Located near the waterfront, its ornate facade is regarded by art historians as one of the world's finest examples of Italian Baroque design. The restoration work left the cathedral looking much as it did when originally completed.

The expansive Plaza de la Revolución, west of Old Havana, is the site of President Castro's major speeches, which are delivered before crowds of, it is estimated, up to 1,000,000 citizens. The plaza is distinguished by some of the city's most imposing architecture. Surrounding the towering monument to José Martí, leader of Cuban independence, are such modern structures as the National Government Centre, the headquarters of the Communist Party of Cuba and the armed forces, and various government ministries. In Central Havana are more traditional buildings, including the white-domed former National Capitol, now housing the Cuban Academy of Sciences; the Museum of the Revolution, housed in the old Presidential Palace; and the National Museum of Art.

The Plaza de la Revolución

Another restoration project was centred on the old Spanish fortifications that dominate Havana's harbour and, for a time in the 17th and 18th centuries, made Havana the most fortified city in Spanish America. The most famous and impressive of these is Morro Castle (Castillo del Morro), completed in 1640. It became the centre of the network of forts protecting Havana, and, with La Punta Fortress (Castillo de la Punta), dominated the actual entrance to the harbour. The oldest fortification, La Fuerza (Castillo de la Fuerza), was begun in 1565 and completed in 1583. Its site at the Plaza de Armas was that of an even older fort erected by Hernando de Soto in 1538 and later destroyed by French pirates.

Morro Castle

THE PEOPLE

Havana, like the rest of Cuba, is populated mostly by people of Spanish ancestry, with a large minority of blacks and mulattoes, whose ancestors were slaves. There are few mestizos, as in many other Latin-American countries, because the Indian population was virtually wiped out in colonial times. In the era before Fidel Castro came to power, the city was economically and ethnically divided. On the one hand, there was the minority of the wealthy, educated elite, together with a developing and expanding middle class, and on the other was the working-class majority. This division was largely based on ethnic background: whites tended to be more well-to-do, while blacks and mulattoes generally were poor. The economic structure did not provide much opportunity for blacks and mulattoes except in the more menial occupations. There was also little opportunity for them to obtain an education.

Pre-Castro ethnic divisions

Under the Castro government that came to power in 1959, this system changed. Educational and employment opportunities were made available to Cubans of all ethnic backgrounds. In housing, the government follows an official policy of no discrimination based on ethnic background, and independent observers tend to believe this policy has been more or less faithfully carried out. Where there were few black or mulatto Cubans in middle- to top-level national and local government posts before 1959, there are now many, although still not in the same proportion as in the population. However, many blacks and mulattoes throughout the island are still in a struggle to lift themselves out of poverty.

Habaneros, as Cubans in general, do not constitute a strongly religious community; about half do not profess a religious affiliation. A number of churches in Havana have continued to operate since the Castro revolution. Roman Catholics form the largest religious group, but the number of parishioners worshiping on a given Sunday is relatively small. There are few priests, and services are

Religion

Alfred Padula



The Cuban Academy of Sciences, formerly the National Capitol, in the Central Havana district.

often conducted by lay people. Protestant churches are similarly limited in activity. The Jewish community in Havana is reduced to only a few hundred from once having embraced more than 50,000 people, many of whom had fled Nazi persecution and subsequently left Cuba after Castro came to power.

One of the major phenomena of the Castro era has been the steady flow of Cubans into exile, mainly to the United States but also to Mexico, Venezuela, and elsewhere in Latin America. In the United States the largest concentration of exiles is in Florida, where they have become a significant minority in the state. It is estimated that at least a million Cubans have left the island since 1959, about 60 percent of whom were said to be Habaneros. Heavy emigrations following the revolution created a shortage of professional people in Havana. This was particularly evident in medicine, law, and economics until well into the 1970s. The Cuban government was forced to give high priority to the training of doctors, dentists, lawyers, and economists to replace these professionals.

Emigration after Castro

THE ECONOMY

Havana's economy first developed on the basis of its location, which made it one of the early great trade centres in the New World. Sugar and a flourishing slave trade first brought riches to the city, and later, after independence, it became a renowned resort. Castro, however, de-emphasized the city as a tourist Mecca.

Industry. Despite efforts by the Castro government to spread Cuba's industrial activity to all parts of the island, Havana remains the centre of much of the nation's industry. The traditional sugar industry, upon which the island's economy has been based for three centuries, is centred elsewhere on the island and controls some three-fourths of the export economy. But light manufacturing facilities, meat-packing plants, and chemical and pharmaceutical operations are concentrated in Havana. Other food-processing industries are also important, along with shipbuilding, vehicle manufacturing, production of alcoholic beverages (particularly rum), textiles, and tobacco products, particularly the world-famous Havana cigars. Although the ports of Cienfuegos and Matanzas, in particular, have been developed under the Castro government, Havana remains Cuba's primary port facility; a majority of Cuban imports and exports pass through Havana. The port also supports a considerable fishing industry.

The sugar industry

Commerce and finance. Under the Castro government, Cuba's traditional free-enterprise system was replaced by a heavily socialized economic system. Most business in Cuba is in the hands of the state. In Havana the large Cuban-owned department stores and U.S.-owned businesses were nationalized and today operate under state control. In Old Havana and throughout Vedado there are a few small private businesses, such as shoe-repair shops or dressmaking facilities, but their number is steadily declining.

Banking has come totally under state control, and the National Bank of Cuba, headquartered in Havana, is the control centre of the Cuban economy. Its branches in some cases occupy buildings that were in pre-Castro times the offices of Cuban or foreign banks.

Transportation. Havana has historically been the hub of Cuba's transportation system and remains so today. This position resulted from Havana's role as the seat of Spanish colonial government and the importance of the port for trade. As railroads developed and as the use of automobiles led to highway construction, the importance of Havana's focal situation increased. Havana became the key terminus for both rail and road links from the east and west. Also, Havana became the main gateway for international air transport. The old Rancho Boyeros airport, now José Martí International Airport, is located eight miles (13 kilometres) from downtown Havana and handles domestic and international flights. A network of bus routes also centres on Havana, and buses are the main mode of inner-city transportation.

Road and railroad development

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Havana is administered by a city council, with a mayor as chief administrative officer. The city is dependent upon the national government, however, for much of its budgetary and overall political direction. The national government is headquartered in Havana and plays an extremely visible role in the city's life. Moreover, the all-embracing authority of many national institutions, including the Communist Party of Cuba (Partido Comunista de Cuba; PCC), the Cuban Army, the militia, and neighbourhood groups called the Committees for the Defense of the Revolution (CDRs), has led to a declining role for the city government, which, nevertheless, still provides such essential services as garbage collection and fire protection. The CDRs, which exist in virtually every street and apartment block, have two main functions: first, to actually defend the revolution against both external and internal opposition and, second, to handle routine tasks in maintaining neighbourhoods.

Role of the CDRs

Havana's borders are contiguous with the province Ciudad de la Habana. Thus Havana functions as both a city and a province. There are two joint councils upon which city and provincial authorities meet—one embraces municipal and provincial leaders on a national basis, the other, a Havana city and provincial council. Havana is divided into 15 constituent municipalities. Until 1976 there were six subdivisions, but in that year the city's borders were expanded to include the entire metropolitan area.

Services. Utility services are under the control of several nationalized state enterprises that have developed since the Castro revolution. Water, electricity, and sewage service are administered in this fashion. Electricity is supplied by generators that are fueled by oil obtained largely from the Soviet Union. Much of the original power plant installation, which operated before the Castro government assumed control, has become somewhat outdated. Electrical blackouts occurred, prompting the national government in 1986 to allocate the equivalent of $25,000,000 to modernize the electrical system. It is said that any part of Havana is within five minutes of a fire-fighting unit; the equipment is largely new, much of it from western Europe and the Soviet Union.

Electric power

Health. Under the Cuban government, all citizens are covered by the national health care plan. Administration of the health care system for the nation is centred largely in Havana. Hospitals in Havana are run by the national government, and citizens are assigned hospitals and clinics to which they may go for attention. During the 1980s Cuba began to attract worldwide attention for its treatment of heart diseases and eye problems, with some of this treatment administered in Havana. There has long been a high standard of health care in the city.

Health care quality

Education. The national government assumes all responsibility for education, and there are adequate primary, secondary, and vocational training schools throughout Havana. The government claims that all children receive an education, and the claim appears valid. The schools are of varying quality, however, and the content of education is clearly aimed at supporting the socialist political orientation of the Castro government. Education is free and compulsory at all levels except higher learning, which is free nonetheless. The University of Havana, located in the Vedado section of Havana, was established in 1728 and was once regarded as a leading institution of higher learning in the Western Hemisphere. Soon after Castro came to power in 1959, the university lost its traditional autonomy and was placed under the control of the government. The city's only other university, the respected Catholic University in Marianao, was closed after the revolution.

Political basis of education

CULTURAL LIFE

Havana, by far the leading cultural centre of the island, offers a wide variety of features that range from museums to ballet and from art and musical festivals to exhibitions of technology. The restoration of Old Havana offered a number of new attractions, including a museum to house relics of the Castro revolution. The government placed special emphasis on cultural activities, many of which are free or involve only a minimal charge.

The Museum of the City of Havana, formerly the Palace of the Captains General in Old Havana, contains many pieces of old furniture, pottery, jewelry, and other exam-

The Museum of the City of Havana

ples of colonial workmanship, as well as models of what Havana looked like in earlier centuries. The museum also houses material relating to the era of U.S. occupation and influence in Cuba. Other important museums are the National Museum of Art in Central Havana and the Museum of Decorative Arts in Vedado. The city's National Library houses Cuba's largest collection. The widest-circulating daily newspapers are published in Havana, but all of these, including the main one, *Granma,* represent Communist Party or government interests.

Dance

The National Ballet of Cuba, led by the world-renowned ballerina Alicia Alonso, performs regularly in Havana, and many of its performances are aimed at bringing dance to schoolchildren. The Cuban ballet performs on state occasions, often in the presence of official state visitors, particularly from socialist-bloc countries.

Many of the city's finest restaurants are in Old Havana. The most popular is Bodeguita del Medio, once a hangout of Ernest Hemingway and now government-run; it serves excellent food despite the rationing that affects most Cubans. La Floridita, also renowned for its Hemingway associations, claims to be the "birthplace of the daiquiri."

Havana was known in pre-Castro times as the queen city of the Caribbean because of its nightlife and popular culture. Much of that has disappeared, but there is still some nightlife, particularly at the fabled Tropicana nightclub. Its present-day dancers and singers are as gaudily and scantily attired as their predecessors were in pre-Castro times, and the stage settings are as big and as imaginative as before. But the songs now often have an anti-American theme, and the audiences are often composed of visiting eastern Europeans rather than Americans.

Carnival

At Carnival time in July, Cubans express themselves vigorously in dance and song. In Havana, Carnival is now a virtual holiday, with floats and parades officially sanctioned. These floats compete in what has become an annual parade along the waterfront Malecón.

Many Cubans are avid sports fans who particularly favour baseball and soccer. Cubans are encouraged to play various sports, and participation is required in the schools.

Baseball

President Castro, once a pitching prospect for the U.S. major leagues, plays baseball from time to time, and his publicity of the game has enhanced its appeal. Habaneros support a dozen or so baseball teams. The city has several large stadiums for soccer and other events. Admission to sporting events is generally free. Social clubs at the beaches provide facilities for water sports and include restaurants and dance halls.

History

FOUNDATION AND EARLY GROWTH

Havana was founded in 1515 by the Spanish conquistador Diego Velázquez de Cuéllar, possibly near the site where the town of Batabanó now stands on the south coast. It was not a fortuitous choice, for the climate was poor and the region was swampy. Mosquitoes abounded. The site was abandoned in favour of the city's present location (then called Puerto Carenas) on the north coast in 1519. The natural deepwater port, together with the land protection to the harbour, made Havana a site that early attracted growing numbers of settlers. A royal decree in 1634 recognized its importance, calling it the "Llave del Nuevo Mundo y Antemural de las Indias Occidentales" ("Key to the New World and Rampart of the West Indies"). Havana's coat of arms carries this inscription.

The importance of harbour fortifications was early recognized as English, French, and Dutch sea marauders attacked the city in the 16th century. The Spaniards began building fortifications, and at the end of the 16th century they transferred the governor's residence to Havana from

The new capital

Santiago de Cuba on the eastern end of the island, thus making Havana the de facto capital in 1592.

DEVELOPMENT AS A MAJOR NEW WORLD PORT

During the 17th century east-bound fleets of Spanish ships carrying treasure from the New World rendezvoused at Havana for the trip across the Atlantic to Spain. The port thus became the object of attacks by competing foreign powers and was blockaded several times during that century. By about 1700 the city walls and the major fortifications had been completed. These withstood attacks until, after a three-month siege ending in August 1762, the British under Admiral Sir George Pocock and the Earl of Albermarle took the city as a prize of war. They held it for six months until the treaty ending the Seven Years' War restored Havana to Spain.

That occupation, as onerous as it was to Habaneros, actually stimulated trade between the New World and Europe, and Havana gained new importance as a port, thriving on the sugar and slave trades. Obstacles to commerce by foreigners were gradually removed as the 18th century ended, and all of Cuba, but Havana in particular, began attracting immigrants from countries other than Spain. This, in turn, added new strains to the ethnic mix of the city—French craftsmen, British merchants, German bankers, and others—and gave Havana a distinct international and cosmopolitan character. Wealthier Cuban colonists visited New York City and Philadelphia. A number of U.S. flagships made port calls at Havana, and there was a small but important U.S. trading community established in Havana by 1850. Nevertheless Cuba remained a Spanish colony despite the wars of independence that raged on the continent in the early 1800s, wars which led to freedom for most of Spain's New World empire.

Growth of U.S. trade

ALLIANCE WITH THE UNITED STATES

Havana, meanwhile, continued to grow as a major world port, rivaling in population and in trade both New York City and Buenos Aires. Cuba became independent from Spain in 1898 with the aid of the United States, and for six decades thereafter Cuba was a close economic and political ally of that country. Increasing numbers of U.S. businesses and tourists were drawn to Havana. Havana acquired the look of a U.S. city with the signs of U.S. economic influence virtually everywhere.

Cuba's government through these years was frequently a dictatorship; at best, it was a fragile democracy, with corruption often running rampant. Many Cubans were unhappy both with the U.S. influence and with the continued dictatorship. There were a number of abortive coup attempts against the government of Fulgencio Batista, but change did not come until the revolution led by Fidel Castro, who on Jan. 1, 1959, took control of Cuba.

Castro takes power

THE CITY UNDER CASTRO

Havana's close ties with the United States were quickly ended, and Castro turned to the Soviet Union for economic and military assistance. Evidence of the Soviet connection thereafter became highly visible in Havana. In the port there were usually a number of Soviet cargo vessels loading or unloading goods; Soviet-made vehicles, both automobiles and trucks, became common in the streets; and Soviet or Soviet-bloc goods appeared in stores.

Despite governmental change and Soviet influence, Havana still retains many of its older traditions, particularly in the narrow streets of Old Havana where past and present merge. This older part of the city deteriorated somewhat after the revolution as Castro first directed resources to the hinterlands. But revival began in the 1980s as restoration and construction plans designed to continue into the 21st century were implemented.

Restoration in the 1980s

BIBLIOGRAPHY. Information on geographic characteristics and on the historical, social, and political development of Havana is found in JAMES D. RUDOLPH (ed.), *Cuba, a Country Study,* 3rd ed. (1985); HUGH THOMAS, *Cuba: The Pursuit of Freedom* (1971); and JORGE I. DOMINGUEZ, *Cuba: Order and Revolution* (1978). Works in Spanish include MARIO VILLAR, "La Habana," pp. 328–419 in vol. 9 of *La enciclopedia de Cuba,* 2nd ed. (1977), a detailed, illustrated essay; and, for the poetic flavour of the city, GASTÓN BAQUERO, "La mítica ciudad llamada La Habana," pp. 420–443 in the same volume. Travel books include PAULA DIPERNA, *The Complete Travel Guide to Cuba* (1979); and GÜNTER GRAU, *Havana: A Complete Travel Guide to the City, Past and Present,* trans. from German (1985). For the history of the city, see JOHN ROBERT MCNEILL, *Atlantic Empires of France and Spain: Louisbourg and Havana, 1700–1763* (1985); and DAVID SYRETT (ed.), *The Siege and Capture of Havana, 1762* (1970). (J.N.G.)

Hebrew Literature

Literature in Hebrew has been produced uninterruptedly from the early 12th century BC, and certain excavated tablets may indicate a literature of even greater antiquity. From 1200 BC to *c.* AD 200, Hebrew was a spoken language in Palestine, first as biblical Hebrew, then as Mishnaic Hebrew, a later dialect that does not derive directly from the biblical dialect and one that gained literary status as the Pharisees began to employ it in their teaching in the 2nd century BC. It was not revived as a spoken language until the late 19th century, and in the 20th century it was adopted as the official language of the new State of Israel. The latter event gave impetus to a growing movement in Hebrew literature centred in Israel.

Hebrew literature is not synonymous with Jewish literature. Some Hebrew writing was produced by the Samaritans and in the 17th century by Protestant enthusiasts. Jews also produced important literatures in Greek, Aramaic, Arabic, Judeo-Spanish (Ladino), Yiddish, and a number of other languages. Apart from the Aramaic writings, however, such literatures always served only that part of Jewry that used the language in question. When that community ceased to exist, the literature produced in languages other than Hebrew was forgotten (or, in the case of Greek Jewish literature, became part of Christian tradition) except for whatever part of it had been translated into Hebrew and thus became part of Hebrew literature.

The Hebrew language, though not spoken between *c.* AD 200 and the late 19th century, has always adapted itself to the needs of changing literary tastes. In the Bible it develops from a simple and earthy idiom to a language suitable for the expression of sophisticated religious thought without losing the poetic force and rhythmic fullness that characterizes it. Mishnaic Hebrew is pedestrian, exact, and yet can reach heights of irony or of warmth. In medieval poetry Hebrew allows extravagant displays of verbal artistry but also, in northwestern Europe, a simplicity equal to that of the spoken languages of its milieu. One generation of translators in the 12th century created a scientific Hebrew that is not inferior to contemporary Arabic or Latin in precision or syntactic refinement. The 17th–19th centuries saw the formation of a stately, rigid, classical style based on biblical Hebrew, but at the same time eastern European mystics made the language serve the expression of their love of God. Literary Hebrew in the 20th century draws upon ancient literature to a marked degree, with styles often modeled upon ancient predecessors. The modern period has also evolved a new type of language for nonliterary writing, while in novels the style is often based upon the spoken language.

This article is divided into the following sections:

Ancient Hebrew literature 483
- Preexilian period, *c.* 1200–587 BC
- Period of the Second Temple, 538 BC–AD 70
- Talmudic literature

Literary revival, 500–1000 484
- Piyyuṭim
- Adoption of Arabic metre

The Middle Ages 484
- The Palestinian tradition in Europe, 800–1300
- The golden age in Spain, 900–1200

The period of retrenchment, 1200–1750 484
- Hebrew culture in western Europe
- Eastern Europe and the religious crisis

The 18th and 19th centuries 485
- Beginnings of the Haskala movement
- Romanticism

Modern literature in Hebrew 485
- Formative influences
- Émigré and Palestinian literature
- Israeli literature

Bibliography 486

ANCIENT HEBREW LITERATURE

Preexilian period, c. 1200–587 BC. All that is preserved of the literature of this period is slightly more than 20 of the 39 books included in the Old Testament (the remainder being from the next period). Poetry probably preceded prose. Biblical poetry was based on the principle of parallelism; *i.e.*, the two halves of a verse express the same idea, either by repeating it in different words or by stressing different aspects of it. Examples are found in the book of Psalms: "But they flattered him with their mouths; they lied to him with their tongues" (Ps. 78:36); "He turned their rivers to blood, so that they could not drink of their streams" (Ps. 78:44). To this form was added a simple rhythm, consisting mainly in having each half of a line divided into an equal number of stressed words. There were also folk songs, to which belonged perhaps large parts of the Song of Solomon, dirges, epic chants, and psalms. The use of various forms of poetry in the work of the prophets appears to be a later development.

The principle of parallelism in biblical poetry

The earlier prose texts were still very close to poetry in structure and language. The first real prose may well have been some of the laws recorded in the Pentateuch. In Jeremiah and Deuteronomy a high standard of prose rhetoric was achieved: some of the conversations in the historical books were attempts to reproduce in writing the style of ordinary speech. (See also BIBLICAL LITERATURE.)

Period of the Second Temple, 538 BC–AD 70. The literary output of this period was large, only part of it belonging to the biblical canon. The biblical Hebrew of the writings was artificial because it had ceased to be spoken and had been replaced by Aramaic, a related Semitic language, and Mishnaic Hebrew. Works that are included among the Dead Sea Scrolls belong to this period. Some of these works provide evidence of a new kind of writing, the homiletic, or sermonizing, commentary to the Bible called Midrash. The only work of real literary merit among the scrolls is the fervent personal poetry of the Hymns of Thanksgiving.

Parts of the biblical books of Ezra and Daniel and certain works among the Dead Sea Scrolls are in an early form of Aramaic. This period also began to provide translations (called Targums) of most of the Hebrew Bible into a slightly later Aramaic.

Talmudic literature. In contrast to the works of the Bible and the Second Temple were the collections of writings concerned with Jewish civil and religious law. Whereas the former were lengthy writings bearing the imprint of their authors or editors, early rabbinic literature consisted entirely of collections of individual statements loosely strung together. The individual paragraphs exhibit the influence of Hellenistic rhetoric. Collections that follow the arrangement of biblical books are called Midrash, as opposed to works such as the Mishna, where the material is arranged according to subject. The Mishna was the main work of the period *c.* 100 BC–AD 200. The following period, AD 200–500, was notable for two main innovations: the appearance of an additional literary centre in Babylonia, where Jewry flourished in contrast to its subjugation under the oppressive rule of Rome and, later, Byzantium in Palestine; and the literary use of the spoken local dialects of Aramaic alongside Hebrew. The Talmuds produced by Palestine and Babylonia in this period contained a large

The Mishna and Midrash

proportion of Haggada, statements dealing with theological and ethical matters and using stories, anecdotes, and parables to illustrate certain points. This material was later an influence on Hebrew fiction of the Middle Ages and of the modern period. (See also JUDAISM: *The literature of Judaism.*)

LITERARY REVIVAL, 500–1000

In the 6th century, some Jewish groups attempted to enforce the exclusive use of Hebrew in the synagogue, this tendency being part of a Hebrew revival that began in Palestine and spread westward but did not reach Babylonia until the 10th century.

Piyyuṭim. Synagogues began in this period to appoint official precentors, part of whose duty it was to compose poetical additions to the liturgy on special sabbaths and festivals. The authors were called *payṭanim* (from Greek *poiētēs,* "poet"), their poems *piyyuṭim.* The keynote was messianic fervour and religious exuberance. Besides employing the entire biblical, Mishnaic, and Aramaic vocabularies, the *payṭanim* coined thousands of new words. Such poems, presupposing a highly educated audience, abound in recondite allusions and contain exhaustive lists of rites and laws. It is known that the most outstanding poets—Phineas the Priest, Yose ben Yose, Yannai, and Eleazar ha-Kalir, or ben Kalir—lived in that order, but when or where in Palestine any of them lived is not known. The accepted datings are 3rd century and 5th–6th century AD. Many *piyyuṭim* are still used in the synagogue.

Adoption of Arabic metre. Biblical Hebrew was reestablished as the literary idiom about 900 by Sa'adia ben Joseph, grammarian and religious polemicist. The Arabic system of quantitative metre was adapted for Hebrew during this period (900–1000), probably by Dunash ben Labrat. At first the *piyyuṭ* form was retained for religious poems, and the new metres were used only for secular poetry, which closely imitated Arabic models and, like the latter, was chiefly employed for laudatory addresses to prominent people.

THE MIDDLE AGES

The Palestinian tradition in Europe, 800–1300. From Palestine, the Hebrew renaissance soon spread into the Byzantine Empire. In Sicily and southern Italy (which belonged to Byzantium) several important *payṭanim* were at work, and before 1000 a secular literature began to arise in Italy: a fantastic travelogue of Eldad the Danite; a historical romance, *Sefer ha-yashar* (1625; Eng. trans., *Sefer ha-yashar, the Book of the Righteous*) and *Josippon,* a revision of Josephus' *Antiquities* filled with legendary incidents—this last-named book was popular until modern times and was translated into many languages. Nathan ben Yehiel completed in 1101 at Rome a dictionary of Talmudic Aramaic and Hebrew, the *'Arukh,* which is still used.

In the middle of the 10th century members of the north Italian family Kalonymos brought Talmudic studies and *piyyuṭim* to Mainz, Ger., where the yeshiva (school) became a centre of studies under the direction of Gershom ben Judah, known as "the Light of the Exile." As a poet, he established a distinctive style of European *piyyuṭ* in poems that read very much like early European popular poetry. The greatest alumnus of the Mainz academy was Rashi, an author of complete commentaries on the Bible and on the Babylonian Talmud, himself a poet of note.

The slaughter of Jewish peoples in western and central Europe during the Crusades drove large masses of Jews into eastern Europe. The German Jews carried with them their Yiddish speech but hardly any literary culture. In Germany accounts of the disaster were written in a new prose style permeated with poetry; liturgical poetry became henceforth mainly a chronicle of persecutions. These sufferings inspired an important mystical movement, largely propagated through stories, of which the chief collections are the *Ayn Shoyn Mayse Bukh* (1602; *Ma'aseh Book*) and the *Sefer Ḥasidim* (1538; "The Book of the Just"), the latter attributed to Judah ben Samuel, "the Hasid" of Regensburg (died 1217).

The golden age in Spain, 900–1200. Spanish Jewry began to flourish in Muslim Spain under the caliphate of Córdoba, where Hasdai ibn Shaprut, a vizier, was the first great patron of Hebrew letters. His secretary, Menahem ben Saruk (died *c.* 970), wrote a biblical lexicon, which was criticized by Dunash ben Labrat when the latter arrived in Spain with philological ideas from the East. Samuel ha-Nagid, vizier of Granada (990–1055), himself a poet and philologist, gathered around him a group of poets, most outstanding among whom was Ibn Gabirol. Moses ibn Ezra of Granada (died *c.* 1139) was the centre of a brilliant circle of poets. Moses' kinsman Abraham ibn Ezra, a poet, philosopher, grammarian, and Bible commentator, attacked the language and style of the early *payṭanim;* he and Judah ben Samuel Halevi were the first to use Arabic metres in religious poems. Dominated by Arab standards of taste, the secular poetry dealt with themes of Arabic poetry and often reproduced Arabic phrases; it was written to be appreciated by a small circle of connoisseurs and declined with the collapse of Jewish prosperity in Muslim Spain. The last major poet in Spain was Judah ben Solomon Harizi, who translated various philosophical works into Hebrew. The poets of Granada

The use of biblical Hebrew was made possible by the work of philologists. Of great importance was the creation of comparative linguistics by Judah ibn Kuraish (about 900) and Isaac ibn Barun (about 1100). Judah Hayyuj, a disciple of Menahem ben Saruk, recast Hebrew grammar, and, in the form given to it by David Kimhi of Narbonne (died *c.* 1235), the new system was taken over by the Christian humanists and through them by modern scholarship. The first complete Hebrew grammar, *Kitāb al-luma'* (1886; "The Book of the Variegated Flower Beds"), was written by Ibn Janāḥ of Córdoba (died 1050).

Jewish medieval philosophers in Spain wrote in Arabic, not Hebrew, until the 13th century. Apart from Isaac Israeli (north Africa, died *c.* 940) few medieval Jews made original contributions to science, but the Spanish Jews shared the best scientific education. Abraham bar Hiyya (died *c.* 1136) of Barcelona was an original mathematician who wrote in Hebrew works on mathematics, astronomy, and philosophy. When the Almohads expelled the Jews from Muslim Spain in 1148, many learned refugees went to Languedoc and Provence and there translated scientific and philosophical works.

THE PERIOD OF RETRENCHMENT, 1200–1750

Hebrew culture in western Europe. From 1200 to 1750 was the era of the ghetto, during which the area of western European Hebrew culture shrank to a remnant in Italy, while an entirely different culture arose in eastern Europe. The appearance in 1200 of the Hebrew version, translated from Arabic, of Moses Maimonides' *Moreh Nevukhim* (1851–85; *The Guide of the Perplexed*), which applied Neoplatonic and Aristotelian philosophy to biblical and rabbinic theology, provoked orthodox circles into opposition to all secular studies. As a result of Maimonides' work, there was a return to Neoplatonist mysticism in a form known as Kabbala. This culminated in the theosophy of the *Zohar* (1560; "The Book of Splendor"), which is ascribed to Moses de Leon and which exercised an influence comparable only with that of the Bible and Talmud. Hebrew culture, however, was reduced to a miniature scale in the West after the expulsion of the Jews from England (1290), from France (1306), and from Spain (1492). It continued in Italy, where it remained in contact with contemporary Christian thought. The most outstanding figure was the mystical philosopher Moshe Ḥayyim Luzzatto, who wrote a work on poetics and three remarkably modern plays. Kabbala

Eastern Europe and the religious crisis. In the kingdom of Poland (which then extended from Lithuania to the Black Sea) refugees from German persecution mingled with earlier Byzantine émigrés to create, by the 15th century, a prosperous Jewry with extensive autonomy. Their culture was not a continuation of western European Hebrew civilization but a new creation. The Bible (except for the Pentateuch) was neglected, while the Babylonian Talmud—hitherto studied only by specialists—became the basis of all intellectual life, particularly since the so-called

pilpul method of Jacob Pollak had turned its study into an exciting form of mental gymnastics. The typical literature consisted of *novellae* (*hiddushim*), ingenious discussions of Talmudic minutiae written in an ungrammatical mixture of Hebrew and Aramaic. Imaginative literature existed only in Yiddish, for women and the uneducated.

The expulsion from Spain produced a wave of messianic emotion. Kabbala flourished in Safad, the new Palestinian centre, the meeting place of Spanish, European, and Oriental Jews. There, in 1570–72, Isaac Luria created a cosmic messianism. Though its formulation, in the writings of his pupil Ḥayyim Vital, was abstruse and esoteric, its phraseology penetrated the widest masses, as a result of the introduction of Kabbalist prayers, and coloured all later Hebrew writing. Luria's teachings were developed by the false messiah Sabbatai Zebi in the next century, for and against whom a vast literature was written.

The sufferings of Polish Jewry in the Cossack massacres of 1648—described in a long poem by the Talmudist Yom Ṭov Lipmann Heller—opened their country to Lurianic mysticism. Out of popular Kabbalist elements, Israel ben Eliezer, called the Ba'al Shem Ṭov, produced Ḥasidism. His teaching, like that of his successors, was oral and, of course, in Yiddish; but it was noted by disciples in a simple, colloquially flavoured Hebrew. Since they taught mainly through parables, this may be considered to mark the beginning of the Hebrew short story. Indeed these narratives exercised, and still exercise, a profound influence on modern Hebrew writers.

Ḥasidism

THE 18TH AND 19TH CENTURIES

In the 18th century the conservative mystical movement of Ḥasidism spread rapidly over all eastern Europe except Lithuania. There, Elijah ben Solomon of Vilna, a writer of unusually wide scope, advocated a better graded course of Talmudic training. Shneur Zalman of Ladi created the highly systematized Ḥabad Ḥasidism, which was widely accepted in Lithuania. The Musar movement of Israel Salanter encouraged the study of medieval ethical writers.

Beginnings of the Haskala movement. In the Berlin of Frederick II the Great, young intellectuals from Poland and elsewhere, brought in as teachers, met representatives of the European Enlightenment; they came under the influence of Moses Mendelssohn and also met some representatives of Italian and Dutch Hebrew cultures. One, a Dane, Naphtali Herz Wessely, who had spent some time in Amsterdam, wrote works on the Hebrew language, and another, an Italian, Samuel Aaron Romanelli, wrote and translated plays. Out of these contacts grew Haskala ("Enlightenment"), a tendency toward westernization that venerated Hebrew and medieval western Jewish literature. Among German Jews, then already in rapid process of Germanization, this Hebrew movement had no place. The Enlightenment was introduced in Galicia (Austrian Poland), a centre of Ḥasidism, by the Edict of Toleration (1781) of the emperor Joseph II. By supporting some of its aims, Hebrew writers incurred hatred and persecution. Their chief weapon was satire, and the imitation by Joseph Perl of the *Epistolae obscurorum virorum* (1515; "Letters of Obscure Men") of Crotus Rubianus and the essays of Isaac Erter were classics of the genre. One poet, Meir Letteris, and one dramatist, Naḥman Isaac Fischman, wrote biblical plays.

Romanticism. Galicia's chief contribution was to the Jüdische Wissenschaft, a school of historical research with Romanticist leanings. The impact of Haskala ideas upon the humanistic Italo-Hebrew tradition produced a short literary renaissance. Its main connections were with the Jüdische Wissenschaft, to which Isaac Samuel Reggio contributed. Samuel David Luzzatto, a prolific essayist, philologist, poet, and letter writer, became prominent by his philosophy of Judaism, while a poet, Rachel Morpurgo, struck some remarkably modern chords. For the Jews of the Russian Empire, the Enlightenment proper began with Isaac Baer Levinsohn in the Ukraine and with Mordecai Aaron Ginzberg (Günzburg), in Lithuania. In the 1820s an orthodox reaction set in, coinciding with the rise of a Romanticist Hebrew school of writers. A.D. Lebensohn wrote fervent love songs to the Hebrew language, and his son Micah Joseph (Mikhal), the most gifted poet of the Haskala period, wrote biblical romances and pantheistic nature lyrics. The first Hebrew novel, *Ahavat Ziyyon* (1853; "The Love of Zion"), by Abraham Mapu, was a Romantic idyll, in which Mapu, like all Haskala writers, employed phrases culled from the Bible and adapted to the thought the writer wished to express.

First Hebrew novel

Mapu's third novel, *'Ayiṭ tzavua'* (1857–69; "The Hypocrite"), marked a departure. It dealt with contemporary life and attacked its social evils and portrayed a new type, the *maskil* (possessor of Haskala), in a fight against orthodox obscurantism. The new, aggressive Haskala soon came under the influence of Russian left-wing writers, such as Nikolay Gavrilovich Chernyshevsky and Dmitry Pisarev. Judah Leib Gordon, like Mapu, had started as a Romantic writer on biblical subjects. From 1871 onward he produced a series of ballads exposing the injustices of traditional Jewish life. Moses Leib Lilienblum began as a moderate religious reformer but later became absorbed by social problems, and in *Mishnat Elisha ben Abuyah* (1878; "The Opinions of Elisha ben Abuyah") he preached Jewish socialism. Peretz Smolenskin created in six novels a kaleidoscope of Jewish life in which he rejected the westernized Jew as much as orthodox reactionaries did.

MODERN LITERATURE IN HEBREW

Formative influences. The first formative influences on 20th-century Hebrew literature belong to the late 19th century. The middle classes of eastern European Jewry that read Hebrew books turned to Jewish nationalism, and Zionist activity, coupled with the movement for speaking Hebrew, widened the circle of Hebrew readers. Hebrew daily papers began to appear in 1886. Writers borrowed extensively from medieval translators and European languages, and the Hebrew language assumed a new character. A key figure in the transition to modern writing was Shalom Jacob Abramowitsch, who wrote under the pseudonym of Mendele Mokher Sefarim; after his first novel he became convinced that biblical Hebrew was unsuitable for modern subjects and turned to Yiddish. From 1886 onward he returned to writing mainly in Hebrew and by using Hebrew and Aramaic phrases from the Talmud was able to capture the homeliness he prized in Yiddish. His stories depicted life as it really was, and his style and support of traditional values attracted a wide readership. The popularity of his stories of ghetto life ensured that they would remain the most read and written genre of Hebrew literature until the mid-20th century. A group of writers adopted "grandfather Mendele" as their model. One of these, Asher Ginzberg (Aḥad Ha'am), wrote, from 1889 onward, articles evolving a secular philosophy of Jewish nationalism. His periodical *ha-Shiloaḥ* attained editorial standards previously unknown in Hebrew. From 1921, he devoted his last years to the editing of his correspondence, a valuable documentary of the period.

Depiction of ghetto life and the trend toward nationalism

Ḥayyim Naḥman Bialik, an important poet, essayist, editor, and anthologist of medieval literature, was, for a time, literary editor of *ha-Shiloaḥ* and was much influenced by Aḥad Ha'am. His poetry expressed the inner struggles of a generation concerned about its attitude to Jewish tradition. Saul Tchernichowsky, on the other hand, was untroubled by tradition, and his poetry dealt with love, beauty, and the three places where he had lived: the Crimea, Germany, and Palestine. Isaac Leib Peretz, who wrote both in Hebrew and in Yiddish, introduced the Ḥasidic, or pietistic devotional, element into literature. The emotionalism and simple joy of life of that milieu thereafter strongly influenced writers, and the language absorbed many Ḥasidic terms. A literary historian, Ruben Brainin, discerned the presence of a "new trend" in literature and foresaw a concentration on human problems. Bialik had already pointed to a conflict between Judaism and the natural instincts of Jews. This psychological interest dominated the work of a group of short-story writers and, in particular, that of the writer and critic David Frischmann, who, more than anyone else, imposed European standards on Hebrew literature. European literary tendencies thus became absorbed into Hebrew. Uprooted by the pogroms of 1881 and the two Russian revolutions of 1905 and 1917,

The conflict between Judaism and the natural instincts of the Jews

Jews had emigrated to western Europe and America, and Hebrew literary activity in eastern Europe was disrupted. The Soviet Union eventually banned Hebrew culture, and it also decayed in other eastern European countries and in Germany as the position of Jews deteriorated.

Émigré and Palestinian literature. The writers of this generation were known as the émigré writers. Their work was pessimistic, as the rootlessness without hope of Uri Nissan Gnessin and Joseph Ḥayyim Brenner exemplified. The majority of writers active in Palestine before 1939 were born in the Diaspora (Jewish communities outside Palestine) and were concerned with the past. An exception was Yehuda Burla, an Oriental Jew who wrote about Oriental Jewry. The transition from ghetto to Palestine was achieved by few writers, among them Asher Barash, who described the early struggles of Palestinian Jewry. Shmuel Yosef Agnon, the outstanding prose writer of this generation (and joint winner of the 1966 Nobel Prize for Literature), developed an original style that borrowed from the Midrash (homiletical commentaries on the Hebrew Scriptures), stories, and ethical writings of earlier centuries. Whereas his earlier stories were set in Galicia, he began in the 1940s to write about Palestine.

In contrast, poetry immediately addressed Palestinian life. Among outstanding writers were Rachel (Rachel Bluwstein), who wrote intensely personal poems; Uri Zevi Greenberg, a political poet and exponent of free verse; and Abraham Shlonsky, who led the Symbolist school.

Israeli literature. World War II and the Arab–Israeli War of 1948–49 brought to the fore Palestinian-born writers who dealt with the problems of their generation in colloquially flavoured Hebrew. In the State of Israel, where Hebrew had become the official language, literature developed on a large scale, mainly along contemporary western European and American lines. The extreme diversity in culture of parts of the population, and the problems of new immigrants, provided the main theme for fiction. Poetry flourished, but original drama at first was slow to develop. Greenberg's *Rehovot HaNahar* (1951; "Streets of the River") traced the process by which the humiliation of the massacred is transmuted by the pride of martyrdom into the historical impulse of messianic redemption. In a long dramatic poem, "Bein ha-Esh ve-ha-Yesha" (1957; *Between the Fire and Salvation*), Aaron Zeitlin envisioned the annihilation of European Jewry in mystical terms, examining the relationship of catastrophe and redemption.

Native Israeli prose writers wrote of their life in the kibbutz, the underground, and the war of 1948–49. S. Yizhar and Moshe Shamir emerged as the outstanding representatives of this generation, probing the sensibility of the individual in a group-oriented society. But the establishment of the State of Israel could not allay the anxieties of the individual. The dominant themes of writers who had no access to collective ideals were personal ones—frustration, confusion, and alienation. The poetry of this period and of following decades, represented, among others, by Yehuda Amichai and Haim Gouri, also emphasizes the dissolution of social coherence and expresses the individual devoid of a sense of historical and spiritual mission. The novelist Aharon Megged's *ha-Hai ʿal ha-met* (1965; *The Living on the Dead*) casts a putative hero of the pioneer generation in an ironic light.

Memories of the Holocaust haunt the lyrical work of Aharon Appelfeld. Flight and hiding are the characteristic situations of his early stories. His *Badenhaim, ʿir nofesh* (*Badenheim 1939*), published in 1975, captures the ominous atmosphere of the approaching Holocaust sensed by a group of assimilated Jews vacationing at an Austrian resort. It describes social and spiritual disintegration as does his novel *Tor ha-peli 'ot* (1978; *The Age of Wonders*). In the stories of Abraham B. Yehoshua the narrator's tone is remote, the people are drained of emotion. Occasionally an act of feeling or meaning breaks the mood of boredom and illuminates a character's humanity. In both *ha-Meʿahev* (1976; *The Lover*) and *Gerushim meʾuḥarim* (1982; *A Late Divorce*) Yehoshua explores the confrontation between the philosophy of the present generation and the ideology of the Zionist founders. Personal frustration and religious vision are the subjects of the novelist Pinḥas Sadeh. Yitzḥak Orpaz's novels tend toward psychological exploration, particularly in the series beginning with *Bayit le-adam eḥad* (1975; "One Man's House"). Yoram Kaniuk examines the alienated Israeli but, in *ha-Yehudi ha-aḥaron* (1981; *The Last Jew*), explores the Israeli experience as a response to the Holocaust. Yitzḥak Ben Ner sets realistic stories in rural and urban communities (*Sheḳiʿah kefarit* [1976; "A Rustic Sunset"] and *Ereẓ reḥokah* [1981; "A Distant Land"]). Yaʿakov Shabtai's novel *Zikhron devarim* (1977; *Past Continuous*) broke new ground in its evocation of the family in society. The world of Amos Oz is generally symbolized by a kibbutz besieged by a powerful and primitive external force always threatening to break in. His *Menuḥah nekhonah* (1982; *A Perfect Peace*) considers the notion of escape from ideological and geographical confinement. Amalia Kahana-Carmon explores the subjective impressions of experience and the complexities of time and memory through a stream-of-consciousness technique.

BIBLIOGRAPHY. JULIUS A. BEWER, *The Literature of the Old Testament*, 3rd ed., completely rev. by EMIL G. KRAELING (1962); SANFORD CALVIN YODER, *Poetry of the Old Testament* (1948, reprinted 1973), a useful introduction and representative collection of biblical poetry; MEYER WAXMAN, *A History of Jewish Literature*, 2nd ed., 5 vol. (1938–60), excellent coverage of Yiddish writing as well as Hebrew; ABRAHAM E. MILLGRAM (ed.), *An Anthology of Medieval Hebrew Literature* (1961); BENZION HALPER, *Post-Biblical Hebrew Literature: An Anthology*, 2 vol. (1921, reprinted 1946); LEON I. FEUER, *Jewish Literature Since the Bible* (1937), with supplements by AZRIEL EISENBERG; T. CARMI (ed.), *The Penguin Book of Hebrew Verse* (1981), a representation of Hebrew poetry from the Bible to the present day, with a useful introduction; MENACHEM RIBALOW, *The Flowering of Modern Hebrew Literature*, trans. from the Hebrew and ed. by JUDAH NADICH (1959); JOSEPH KLAUSNER, *A History of Modern Hebrew Literature, 1788–1930* (1932, reprinted 1972; originally published in Hebrew, 1920); JACOB S. RAISIN, *The Haskalah Movement in Russia* (1913, reprinted 1972); DOV VARDY, *The New Hebrew Poetry* (1947); SAMUELE AVISAR, *Teatro Ebraico* (1957); DAVID PATTERSON, *The Hebrew Novel in Czarist Russia* (1964); MENACHEM RIBALOW, *Anthology of Hebrew Poetry in America* (1938), with text in Hebrew; SIMON HALKIN, *Modern Hebrew Literature, from the Enlightenment to the Birth of the State of Israel: Trends and Values*, new ed. (1970); JACOB J. PETUCHOWSKI, *Theology and Poetry: Studies in the Medieval Piyyut* (1978), with a general introduction, text, and notes; STANLEY BURNSHAW, T. CARMI, and EZRA SPICEHANDLER (eds.), *The Modern Hebrew Poem Itself* (1965), in Hebrew with English transcription and rendering; EISIG SILBERSCHLAG, *From Renaissance to Renaissance*, vol. 2: *Hebrew Literature in the Land of Israel, 1870–1970* (1977); GLENDA ABRAMSON, *Modern Hebrew Drama* (1979); ELLIOTT ANDERSON (ed.), *Contemporary Israeli Literature: An Anthology* (1977); JACOB SONNTAG (ed.), *New Writing from Israel 1976: Stories, Poems, Essays* (1976).

(Ch.R./S.Lr./G.M.A.)

Hegel and Hegelianism

G.W.F. Hegel was the last of the great philosophical-system builders of modern times. His work, following upon that of Immanuel Kant, Johann Gottlieb Fichte, and Friedrich Schelling, thus marks the pinnacle of classical German philosophy. As an absolute Idealist inspired by Christian insights and grounded in his mastery of a fantastic fund of concrete knowledge, Hegel found a place for everything—logical, natural, human, and divine—in a dialectical scheme that repeatedly swung from thesis to antithesis and back again to a higher and richer synthesis. His influence has been as fertile in the reactions that he precipitated—in Søren Kierkegaard, the Danish Existentialist; in the Marxists, who turned to social action; in the Vienna Positivists; and in G.E. Moore, a pioneering figure in British Analytic philosophy—as in his positive impact.

This article deals with the man and his accomplishments as well as with the philosophical movement, Hegelianism, that evolved from his thought. It is divided into the following sections:

Life and work 487
- Early life 487
  - Emancipation from Kantianism
  - Career as lecturer at Jena
- Gymnasium rector 489
- University professor 489
  - At Heidelberg
  - At Berlin
- Personage and influence 490

Hegelianism 491
- General considerations 491
  - Problems of the Hegelian heritage
  - Stages in the history of the interpretation of Hegel
- Crises in the earlier Hegelian school 491
  - Polemics during the life of Hegel: 1816–31
  - Period of controversies chiefly in religion: 1831–39
  - Period of atheistic and political radicalism: 1840–44
- Hegelianism through the 20th century 494
  - Development and diffusion of Hegelianism in the later 19th century
  - Hegelianism in the first half of the 20th century
  - Hegelian studies today

Bibliography 496

Life and work

EARLY LIFE

Hegel—who was born in Stuttgart on August 27, 1770, the son of a revenue officer—was christened Georg Wilhelm Friedrich. He had already learned the elements of Latin from his mother by the time he entered the Stuttgart grammar school, where he remained for his education until he was 18. As a schoolboy he made a collection of extracts, alphabetically arranged, comprising annotations on classical authors, passages from newspapers, and treatises on morals and mathematics from the standard works of the period.

University and private studies

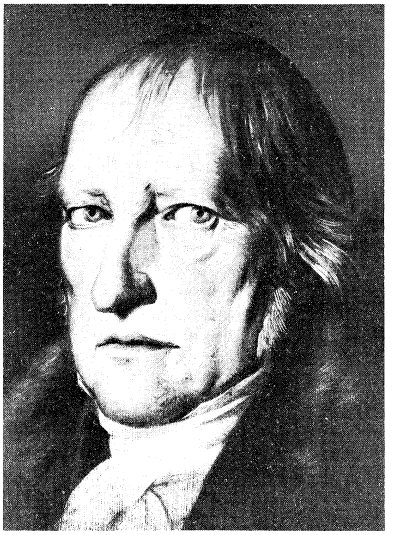
In 1788 Hegel went as a student to Tübingen with a view to taking orders, as his parents wished. Here he studied philosophy and classics for two years and graduated in 1790. Though he then took the theological course, he was impatient with the orthodoxy of his teachers; and the certificate given to him when he left in 1793 states that, whereas he had devoted himself vigorously to philosophy, his industry in theology was intermittent. He was also said to be poor in oral exposition, a deficiency that was to dog him throughout his life. Though his fellow students called him "the old man," he liked cheerful company and a "sacrifice to Bacchus" and enjoyed the ladies as well. His chief friends during that period were a pantheistic poet, J.C.F. Hölderlin, his contemporary, and the nature philosopher Schelling, five years his junior. Together they read the Greek tragedians and celebrated the glories of the French Revolution.

On leaving college, Hegel did not enter the ministry; instead, wishing to have leisure for the study of philosophy and Greek literature, he became a private tutor. For the next three years he lived in Berne, with time on his hands and the run of a good library, where he read Edward Gibbon on the fall of the Roman empire and *De l'esprit des loix,* by Charles Louis, baron de Montesquieu, as well as the Greek and Roman classics. He also studied the critical philosopher Immanuel Kant and was stimulated by his essay on religion to write certain papers that became noteworthy only when, more than a century later, they were published as a part of *Hegels theologische Jugendschriften* (1907). Kant had maintained that, whereas orthodoxy requires a faith in historical facts and in doctrines that reason alone cannot justify and imposes on the faithful a moral system of arbitrary commands alleged to be revealed, Jesus, on the contrary, had originally taught a rational morality, which was reconcilable with the teaching of Kant's ethical works, and a religion that, unlike Judaism, was adapted to the reason of all men. Hegel accepted this teaching; but, being more of a historian than Kant was, he put it to the test of history by writing two essays. The first of these was a life of Jesus in which Hegel attempted to reinterpret the gospel on Kantian lines. The second essay was an answer to the question of how Christianity had ever become the authoritarian religion that it was, if in fact the teaching of Jesus was not authoritarian but rationalistic.

Move to Frankfurt am Main

Hegel was lonely in Berne and was glad to move, at the end of 1796, to Frankfurt am Main, where Hölderlin had gotten him a tutorship. His hopes of more companionship, however, were unfulfilled: Hölderlin was engrossed in an illicit love affair and shortly lost his reason. Hegel began to suffer from melancholia and, to cure himself, worked harder than ever, especially at Greek philosophy and modern history and politics. He read and made clippings from English newspapers, wrote about the internal affairs of his native Wurtemberg, and studied economics. Hegel was now able to free himself from the domination

Deutsche Fotothek, Dresden



Hegel, oil painting by Jakob von Schlesinger, *c.* 1825. In the Staatliche Museen zu Berlin.

of Kant's influence and to look with a fresh eye on the problem of Christian origins.

Emancipation from Kantianism. It is impossible to exaggerate the importance that this problem had for Hegel. It is true that his early theological writings contain hard sayings about Christianity and the churches; but the object of his attack was orthodoxy, not theology itself. All that he wrote at this period throbs with a religious conviction of a kind that is totally absent from Kant and Hegel's other 18th-century teachers. Above all, he was inspired by a doctrine of the Holy Spirit. The spirit of man, his reason, is the candle of the Lord, he held, and therefore cannot be subject to the limitations that Kant had imposed upon it. This faith in reason, with its religious basis, henceforth animated the whole of Hegel's work.

His outlook had also become that of a historian—which again distinguishes him from Kant, who was much more influenced by the concepts of physical science. Every one of Hegel's major works was a history; and, indeed, it was among historians and classical scholars rather than among philosophers that his work mainly fructified in the 19th century.

When in 1798 Hegel turned back to look over the essays that he had written in Berne two or three years earlier, he saw with a historian's eye that, under Kant's influence, he had misrepresented the life and teachings of Jesus and the history of the Christian Church. His newly won insight then found expression in his essay "Der Geist des Christentums und sein Schicksal" ("The Spirit of Christianity and Its Fate"), likewise unpublished until 1907. This is one of Hegel's most remarkable works. Its style is often difficult and the connection of thought not always plain, but it is written with passion, insight, and conviction.

"The Spirit of Christianity"

He begins by sketching the essence of Judaism, which he paints in the darkest colours. The Jews were slaves to the Mosaic Law, leading a life unlovely in comparison with that of the ancient Greeks and content with the material satisfaction of a land flowing with milk and honey. Jesus taught something entirely different. Men are not to be the slaves of objective commands: the law is made for man. They are even to rise above the tension in moral experience between inclination and reason's law of duty, for the law is to be "fulfilled" in the love of God, wherein all tension ceases and the believer does God's will wholeheartedly and single-mindedly. A community of such believers is the Kingdom of God.

This is the kingdom that Jesus came to teach. It is founded on a belief in the unity of the divine and the human. The life that flows in them both is one; and it is only because man is spirit that he can grasp and comprehend the Spirit of God. Hegel works out this conception in an exegesis of passages in the Gospel According to John. The kingdom, however, can never be realized in this world: man is not spirit alone but flesh also. "Church and state, worship and life, piety and virtue, spiritual and worldly action can never dissolve into one."

In this essay the leading ideas of Hegel's system of philosophy are rooted. Kant had argued that man can have knowledge only of a finite world of appearances and that, whenever his reason attempts to go beyond this sphere and grapple with the infinite or with ultimate reality, it becomes entangled in insoluble contradictions. Hegel, however, found in love, conceived as a union of opposites, a prefigurement of spirit as the unity in which contradictions, such as infinite and finite, are embraced and synthesized. His choice of the word *Geist* to express this his leading conception was deliberate: the word means "spirit" as well as "mind" and thus has religious overtones. Contradictions in thinking at the scientific level of Kant's "understanding" are indeed inevitable, but thinking as an activity of spirit or "reason" can rise above them to a synthesis in which the contradictions are resolved. All of this, expressed in religious phraseology, is contained in the manuscripts written toward the end of Hegel's stay in Frankfurt. "In religion," he wrote, "finite life rises to infinite life." Kant's philosophy had to stop short of religion. But there is room for another philosophy, based on the concept of spirit, that will distill into conceptual form the insights of religion. This was the philosophy that Hegel now felt himself ready to expound.

Career as lecturer at Jena. Fortunately, his circumstances changed at this moment, and he was at last able to embark on the academic career that had long been his ambition. His father's death in 1799 had left him an inheritance, slender, indeed, but sufficient to enable him to surrender a regular income and take the risk of becoming a *Privatdozent.* In January of 1801 he arrived in Jena, where Schelling had been a professor since 1798. Jena, which had harboured the fantastic mysticism of the Schlegel brothers and their colleagues and the Kantianism and ethical Idealism of Fichte, had already seen its golden age, for these great scholars had all left. The precocious Schelling, who was but 26 on Hegel's arrival, already had several books to his credit. Apt to "philosophize in public," Schelling had been fighting a lone battle in the university against the rather dull followers of Kant. It was suggested that Hegel had been summoned as a new champion to aid his friend. This impression received some confirmation from the dissertation by which Hegel qualified as a university teacher, which betrays the influence of Schelling's philosophy of nature, as well as from Hegel's first publication, an essay entitled "Differenz des Fichte'schen und Schelling'schen Systems der Philosophie" (1801), in which he gave preference to the latter. Nevertheless, even in this essay and still more in its successors, Hegel's difference from Schelling was clearly marked; they had a common interest in the Greeks, they both wished to carry forward Kant's work, they were both iconoclasts; but Schelling had too many romantic enthusiasms for Hegel's liking; and all that Hegel took from him—and then only for a very short period—was a terminology.

Relationship with Schelling

Hegel's lectures, delivered in the winter of 1801–02, on logic and metaphysics, were attended by about 11 students. Later, in 1804, with a class of about 30, he lectured on his whole system, gradually working it out as he taught. Notice after notice of his lectures promised a textbook of philosophy—which, however, failed to appear. After the departure of Schelling from Jena (1803), Hegel was left to work out his own views untrammelled. Besides philosophical and political studies, he made extracts from books, attended lectures on physiology, and dabbled in other sciences. As a result of representations made by himself at Weimar, he was in February 1805 appointed extraordinary professor at Jena; and in July 1806, on Goethe's intervention, he drew his first stipend—100 thalers. Though some of his hearers became attached to him, Hegel was not yet a popular lecturer.

Hegel, like Goethe, felt no patriotic shudder when Napoleon won his victory at Jena (1806): in Prussia he saw only a corrupt and conceited bureaucracy. Writing to a friend on the day before the battle, he spoke with admiration of the "world soul" and the Emperor and with satisfaction at the probable overthrow of the Prussians.

At this time Hegel published his first great work, the *Phänomenologie des Geistes* (1807; Eng. trans., *The Phenomenology of Mind,* 2nd ed., 1931). This, perhaps the most brilliant and difficult of Hegel's books, describes how the human mind has risen from mere consciousness, through self-consciousness, reason, spirit, and religion, to absolute knowledge. Though man's native attitude toward existence is reliance on the senses, a little reflection is sufficient to show that the reality attributed to the external world is due as much to intellectual conceptions as to the senses and that these conceptions elude a man when he tries to fix them. If consciousness cannot detect a permanent object outside itself, so self-consciousness cannot find a permanent subject in itself. Through aloofness, skepticism, or imperfection, self-consciousness has isolated itself from the world; it has closed its gates against the stream of life. The perception of this is reason. Reason thus abandons its efforts to mold the world and is content to let the aims of individuals work out their results independently.

The Phenomenology of Mind

The stage of *Geist,* however, reveals the consciousness no longer as isolated, critical, and antagonistic but as the indwelling spirit of a community. This is the lowest stage of concrete consciousness, the age of unconscious morality. But, through increasing culture, the mind gradually

emancipates itself from conventions, which prepares the way for the rule of conscience. From the moral world the next step is religion. But the idea of Godhead, too, has to pass through nature worship and art before it reaches a full utterance in Christianity. Religion thus approaches the stage of absolute knowledge, of "the spirit knowing itself as spirit." Here, according to Hegel, is the field of philosophy.

GYMNASIUM RECTOR

In spite of the *Phänomenologie,* however, Hegel's fortunes were now at their lowest ebb. He was, therefore, glad to become editor of the *Bamberger Zeitung* (1807–08). This, however, was not a suitable vocation, and he gladly accepted the rectorship of the Aegidiengymnasium in Nürnberg, a post he held from December 1808 to August 1816 and one that offered him a small but assured income. There Hegel inspired confidence in his pupils and maintained discipline without pedantic interference in their associations and sports.

In 1811 Hegel married Marie von Tucher (22 years his junior), of Nürnberg. The marriage was entirely happy. His wife bore him two sons: Karl, who became eminent as a historian; and Immanuel, whose interests were theological. The family circle was joined by Ludwig, a natural son of Hegel's from Jena. At Nürnberg in 1812 appeared *Die objektive Logik,* being the first part of his *Wissenschaft der Logik* ("Science of Logic"), which in 1816 was completed by the second part, *Die subjecktive Logik.*

UNIVERSITY PROFESSOR

This work, in which his system was first presented in what was essentially its ultimate shape, earned him the offer of professorships at Erlangen, at Berlin, and at Heidelberg.

Hegel's philosophical system

At Heidelberg. He accepted the chair at Heidelberg. For use at his lectures there, he published his *Encyklopädie der philosophischen Wissenschaften im Grundrisse* (1817; "Encyclopaedia of the Philosophical Sciences in Outline"), an exposition of his system as a whole. Hegel's philosophy is an attempt to comprehend the entire universe as a systematic whole. The system is grounded in faith. In the Christian religion God has been revealed as truth and as spirit. As spirit, man can receive this revelation. In religion the truth is veiled in imagery; but in philosophy the veil is torn aside, so that man can know the infinite and see all things in God. Hegel's system is thus a spiritual monism but a monism in which differentiation is essential. Only through an experience of difference can the identity of thought and the object of thought be achieved—an identity in which thinking attains the through-and-through intelligibility that is its goal. Thus, truth is known only because error has been experienced and truth has triumphed; and God is infinite only because he has assumed the limitations of finitude and triumphed over them. Similarly, man's Fall was necessary if he was to attain moral goodness. Spirit, including the Infinite Spirit, knows itself as spirit only by contrast with nature. Hegel's system is monistic in having a single theme: what makes the universe intelligible is to see it as the eternal cyclical process whereby Absolute Spirit comes to knowledge of itself as spirit (1) through its own thinking; (2) through nature; and (3) through finite spirits and their self-expression in history and their self-discovery, in art, in religion, and in philosophy, as one with Absolute Spirit itself.

"Encyclopaedia of the Philosophical Sciences"

The compendium of Hegel's system, the "Encyclopaedia of the Philosophical Sciences," is in three parts: "Logic," "Nature," and "Mind." Hegel's method of exposition is dialectical. It often happens that in a discussion two people who at first present diametrically opposed points of view ultimately agree to reject their own partial views and to accept a new and broader view that does justice to the substance of each. Hegel believed that thinking always proceeds according to this pattern: it begins by laying down a positive thesis that is at once negated by its antithesis; then further thought produces the synthesis. But this in turn generates an antithesis, and the same process continues once more. The process, however, is circular: ultimately, thinking reaches a synthesis that is identical with its starting point, except that all that was implicit there has now been made explicit. Thus, thinking itself, as a process, has negativity as one of its constituent moments, and the finite is, as God's self-manifestation, part and parcel of the infinite itself. This is the sort of dialectical process of which Hegel's system provides an account in three phases.

"Logic." The system begins with an account of God's thinking "before the creation of nature and finite spirit"; *i.e.,* with the categories or pure forms of thought, which are the structure of all physical and intellectual life. Throughout, Hegel is dealing with pure essentialities, with spirit thinking its own essence; and these are linked together in a dialectical process that advances from abstract to concrete. If a man tries to think the notion of pure Being (the most abstract category of all), he finds that it is simply emptiness; *i.e.,* Nothing. Yet Nothing *is.* The notion of pure Being and the notion of Nothing are opposites; and yet each, as one tries to think it, passes over into the other. But the way out of the contradiction is at once to reject both notions separately and to affirm them both together; *i.e.,* to assert the notion of becoming, since what becomes both is and is not at once. The dialectical process advances through categories of increasing complexity and culminates with the absolute idea, or with the spirit as objective to itself.

"Nature." Nature is the opposite of spirit. The categories studied in "Logic" were all internally related to one another; they grew out of one another. Nature, on the other hand, is a sphere of external relations. Parts of space and moments of time exclude one another; and everything in nature is in space and time and is thus finite. But nature is created by spirit and bears the mark of its creator. Categories appear in it as its essential structure, and it is the task of the philosophy of nature to detect that structure and its dialectic; but nature, as the realm of externality, cannot be rational through and through, though the rationality prefigured in it becomes gradually explicit when man appears. In man nature rises to self-consciousness.

"Mind." Here Hegel follows the development of the human mind through the subconscious, consciousness, and the rational will; then through human institutions and human history as the embodiment or objectification of that will; and finally to art, religion, and philosophy, in which finally man knows himself as spirit, as one with God and possessed of absolute truth. Thus, it is now open to him to think his own essence; *i.e.,* the thoughts expounded in "Logic." He has finally returned to the starting point of the system, but en route he has made explicit all that was implicit in it and has discovered that "nothing but spirit is, and spirit is pure activity."

Hegel's system depends throughout on the results of scientific, historical, theological, and philosophical inquiry. No reader can fail to be impressed by the penetration and breadth of his mind nor by the immense range of knowledge that, in his view, had to precede the work of philosophizing. A civilization must be mature and, indeed, in its death throes before, in the philosophic thinking that has implicitly been its substance, it becomes conscious of itself and of its own significance. Thus, when philosophy comes on the scene, some form of the world has grown old.

At Berlin. In 1818 Hegel accepted the renewed offer of the chair of philosophy at Berlin, which had been vacant since Fichte's death. There his influence over his pupils was immense, and there he published his *Naturrecht und Staatswissenschaft im Grundrisse,* alternatively entitled *Grundlinien der Philosophie des Rechts* (1821; Eng. trans., *The Philosophy of Right,* 1942). In Hegel's works on politics and history, the human mind objectifies itself in its endeavour to find an object identical with itself. *The Philosophy of Right* (or *of Law*) falls into three main divisions. The first is concerned with law and rights as such: persons (*i.e.,* men as men, quite independently of their individual characters) are the subject of rights, and what is required of them is mere obedience, no matter what the motives of obedience may be. Right is thus an abstract universal and therefore does justice only to the universal element in the human will. The individual, however, cannot be satisfied unless the act that he does accords not merely with law but also with his own conscientious convictions. Thus, the problem in the modern world is to construct a social

The Philosophy of Right

and political order that satisfies the claims of both. And thus no political order can satisfy the demands of reason unless it is organized so as to avoid, on the one hand, a centralization that would make men slaves or ignore conscience and, on the other hand, an antinomianism that would allow freedom of conviction to any individual and so produce a licentiousness that would make social and political order impossible. The state that achieves this synthesis rests on the family and on the guild. It is unlike any state existing in Hegel's day; it is a form of limited monarchy, with parliamentary government, trial by jury, and toleration for Jews and dissenters.

After his publication of *The Philosophy of Right,* Hegel seems to have devoted himself almost entirely to his lectures. Between 1823 and 1827 his activity reached its maximum. His notes were subjected to perpetual revisions and additions. It is possible to form an idea of them from the shape in which they appear in his published writings. Those on *Aesthetics,* on the *Philosophy of Religion,* on the *Philosophy of History,* and on the *History of Philosophy* have been published by his editors, mainly from the notes of his students, whereas those on logic, psychology, and the philosophy of nature have been appended in the form of illustrative and explanatory notes to the corresponding sections of his *Encyklopädie.* During these years hundreds of hearers from all parts of Germany and beyond came under his influence; and his fame was carried abroad by eager or intelligent disciples.

Philosophies of aesthetics; of religion; of history

Three courses of lectures are especially the product of his Berlin period: those on aesthetics, on the philosophy of religion, and on the philosophy of history. In the years preceding the revolution of 1830, public interest, excluded from political life, turned to theatres, concert rooms, and picture galleries. At these Hegel became a frequent and appreciative visitor, and he made extracts from the art notes in the newspapers. During his holiday excursions, his interest in the fine arts more than once took him out of his way to see some old painting. This familiarity with the facts of art, though neither deep nor historical, gave a freshness to his lectures on aesthetics, which, as put together from the notes taken in different years from 1820 to 1829, are among his most successful efforts.

The lectures on the philosophy of religion are another application of his method, and shortly before his death he had prepared for the press a course of lectures on the proofs for the existence of God. On the one hand, he turned his weapons against the Rationalistic school, which reduced religion to the modicum compatible with an ordinary worldly mind. On the other hand, he criticized the school of Schleiermacher, who elevated feeling to a place in religion above systematic theology. In his middle way, Hegel attempted to show that the dogmatic creed is the rational development of what was implicit in religious feeling. To do so, of course, philosophy must be made the interpreter and the superior discipline.

In his philosophy of history, Hegel presupposed that the whole of human history is a process through which mankind has been making spiritual and moral progress and advancing to self-knowledge. History has a plot, and the philosopher's task is to discern it. Some historians have found its key in the operation of natural laws of various kinds. Hegel's attitude, however, rested on the faith that history is the enactment of God's purpose and that man had now advanced far enough to descry what that purpose is: it is the gradual realization of human freedom.

The first step was to make the transition from a natural life of savagery to a state of order and law. States had to be founded by force and violence; there is no other way to make men law-abiding before they have advanced far enough mentally to accept the rationality of an ordered life. There will be a stage at which some men have accepted the law and become free, while others remain slaves. In the modern world man has come to appreciate that all men, as minds, are free in essence, and his task is thus to frame institutions under which they will be free in fact.

Hegel did not believe, despite the charge of some critics, that history had ended in his lifetime. In particular, he maintained against Kant that to eliminate war is impossible. Each nation-state is an individual; and, as Hobbes had said of relations between individuals in the state of nature, pacts without the sword are but words. Clearly, Hegel's reverence for fact prevented him from accepting Kant's Idealism.

The lectures on the history of philosophy are especially remarkable for their treatment of Greek philosophy. Working without modern indexes and annotated editions, Hegel's grasp of Plato and Aristotle is astounding, and it is only just to recognize that it was from Hegel that the scholarship lavished on Greek philosophy in the century after his death received its original impetus.

Recognition and honours

At this time a Hegelian school began to gather. The flock included intelligent pupils, empty-headed imitators, and romantics who turned philosophy into lyric measures. Opposition and criticism only served to define more precisely the adherents of the new doctrine. Though he had soon resigned all direct official connection with the schools of Brandenburg, Hegel's real influence in Prussia was considerable. In 1830 he was rector of the university. In 1831 he received a decoration from Frederick William III. One of his last literary undertakings was the establishment of the Berlin *Jahrbücher für wissenschaftliche Kritik* ("Yearbook for Philosophical Criticism").

The revolution of 1830 was a great blow to Hegel, and the prospect of mob rule almost made him ill. His last literary work, the first part of which appeared in the *Preussische Staatszeitung* while the rest was censored, was an essay on the English Reform Bill of 1832, considering its probable effects on the character of the new members of Parliament and the measures that they might introduce. In the latter connection he enlarged on several points in which England had done less than many continental states for the abolition of monopolies and abuses.

In 1831 cholera entered Germany. Hegel and his family retired for the summer to the suburbs, and there he finished the revision of the first part of his *Science of Logic.* Home again for the winter session, on November 14, after one day's illness, he died of cholera and was buried, as he had wished, between Fichte and Karl Solger, author of an ironic dialectic.

PERSONAGE AND INFLUENCE

In his classroom Hegel was more impressive than fascinating. His students saw a plain, old-fashioned face, without life or lustre—a figure that had never looked young and was now prematurely aged. Sitting with his snuffbox before him and his head bent down, he looked ill at ease and kept turning the folios of his notes. His utterance was interrupted by frequent coughing; every sentence came out with a struggle. The style was no less irregular: sometimes in plain narrative the lecturer would be specially awkward, while in abstruse passages he seemed especially at home, rose into a natural eloquence, and carried away the hearer by the grandeur of his diction.

The early theological writings and the *Phenomenology of Mind* are packed with brilliant metaphors. In his later works, produced as textbooks for his lectures, the "Encyclopaedia of the Philosophical Sciences" and the *Philosophy of Right,* he compresses his material into relatively short, numbered paragraphs. It is only necessary to translate them to appreciate their conciseness and precision. The common idea that Hegel's is a philosophy of exceptional difficulty is quite mistaken. Once his terminology is understood and his main principles grasped, he presents far less difficulty than Kant, for example. One reason for this is a certain air of dogmatism: Kant's statements are often hedged around with qualifications; but Hegel had, as it were, seen a vision of absolute truth, and he expounds it with confidence.

Emergence of Hegelianism

Hegel's system is avowedly an attempt to unify opposites—spirit and nature, universal and particular, ideal and real—and to be a synthesis in which all the partial and contradictory philosophies of his predecessors are alike contained and transcended. It is thus both Idealism and Realism at once; hence, it is not surprising that his successors, emphasizing now one and now another strain in his thought, have interpreted him variously. Conservatives and revolutionaries, believers and atheists alike have professed to draw inspiration from him. In one form or

another his teaching dominated German universities for some years after his death and spread to France and to Italy. The vicissitudes of Hegelian thought to the present day are detailed below in *Hegelianism.* In the mid-20th century, interest in the early theological writings and in the *Phänomenologie* was increased by the spread of Existentialism. At the same time, the growing importance of Communism encouraged political thinkers to study Hegel's political works, as well as his "Logic," because of their influence on Karl Marx. And, by the time of his bicentennial in 1970, a Hegelian renascence was in the making. (T.M.K.)

Hegelianism

Hegelianism is the name given to a diversified philosophical movement that developed out of Hegel's monumental system of thought. The term is here so construed as to exclude Hegel himself and to include, therefore, only the ensuing Hegelian movements. As such, its thought is focussed upon history and logic, a history in which it sees, in various perspectives, that "the rational is the real" and a logic in which it sees that "the truth is the Whole."

GENERAL CONSIDERATIONS

Problems of the Hegelian heritage. The Hegelian system, in which German Idealism reached its fulfillment, claimed to provide a unitary solution to all of the problems of philosophy. It held that the speculative point of view, which transcends all particular and separate perspectives, must grasp the *one* truth, bringing back to its proper centre all of the problems of logic, of metaphysics (or the nature of Being), and of the philosophies of nature, law, history, and culture (artistic, religious, and philosophical). According to Hegel, this attitude is more than a formal method that remains extraneous to its own content; rather, it represents the actual development of the Absolute—of the all-embracing totality of reality—considered "as Subject and not merely as Substance" (*i.e.,* as a conscious agent or Spirit and not merely as a real being). This Absolute, Hegel held, first puts forth (or posits) itself in the immediacy of its own inner consciousness and then negates this positing—expressing itself now in the particularity and determinateness of the factual elements of life and culture—and finally regains itself, through the negation of the former negation that had constituted the finite world.

Dialectical development of the Absolute

Such a dialectical scheme (immediateness–alienation–negation of the negation) accomplished the self-resolution of the aforementioned problem areas—of logic, of metaphysics, and so on. This panoramic system thus had the merit of engaging philosophy in the consideration of all of the problems of history and culture, none of which could any longer be deemed foreign to its competence. At the same time, however, the system deprived all of the implicated elements and problems of their autonomy and particular authenticity, reducing them to symbolic manifestations of the one process, that of the Absolute Spirit's quest for and conquest of its own self. Moreover, such a speculative mediation between opposites, when directed to the more impending problems of the time, such as those of religion and politics, led ultimately to the evasion of the most urgent and imperious ideological demands and was hardly able to escape the charge of ambiguity and opportunism.

Stages in the history of the interpretation of Hegel. The explanation of the success of Hegelianism—marked by the formation of a school that, for more than 30 years, brought together the best energies of German philosophy—lies in the fact that no other system could compete with it in the richness of its content or the rigour of its formulation or challenge its claim to express the total spirit of the culture of its time. Moreover, as Hegelianism diffused outward, it was destined to provoke increasingly lively and gripping reactions and to take on various articulations as, in its historical development, it intermingled with contrasting positions.

Four stages can be distinguished within the development of Hegelianism. The first of these was that of the immediate crisis of the Hegelian school in Germany during the period from 1827 through 1850. Always involved in polemics against its adversaries, the school soon divided into three currents: (1) the right, in which the direct disciples of Hegel participated, defended his philosophy from the accusation that it was liberal and pantheistic (defining God as the All). These "old Hegelians" sought to uphold the compatibility of Hegelianism with evangelical orthodoxy and with the conservative political policies of the Restoration (the new order in Europe that followed the defeat of Napoleon). (2) The left—formed of the "young Hegelians," for the most part indirect disciples of Hegel—considered the dialectic as a "principle of movement" and viewed Hegel's identification of the rational with the real as a command to modify the cultural and political reality that reactionism was merely justifying and to make it rational. Thus the young Hegelians interpreted Hegelianism in a revolutionary sense—*i.e.,* as pantheistic and then, consecutively, as atheistic in religion and as liberal democratic in politics. (3) The centre, which preferred to fall back upon interpretations of the Hegelian system in its genesis and significance, with special interest in logical problems.

Hegelian right, centre, and left

In the second phase (1850–1904), in which Hegelianism diffused into other countries, the works of the centre played a preponderant role; thus in this phase of the history of the interpretation of Hegel, usually called Neo-Hegelian, the primary interest was in logic and a reform of the dialectic.

In the first decade of the 20th century, on the other hand, there arose still in Germany a different movement, after Wilhelm Dilthey, originator of a critical approach to history and humanistic studies, discovered unpublished papers from the period of Hegel's youth. This third phase, that of the Hegel renaissance, was characterized by an interest in philology, by the publication of texts, and by historical studies; and it stressed the reconstruction of the genesis of Hegel's thought, considering especially its cultural matrices—both Enlightenment and Romanticist—and the extent to which it might present irrationalistic and so-called pre-Existentialist attitudes.

In the fourth stage, after World War II, the revival of Marxist studies in Europe finally thrust into the foreground the interest in Hegel–Marx relationships and in the value of the Hegelian heritage for Marxism, with particular regard to political and social problems. This fourth phase of the history of Hegelianism thus appropriated many of the polemical themes of the earlier years of the school.

CRISES IN THE EARLIER HEGELIAN SCHOOL

The earlier development of Hegelianism can be divided, according to predominant concerns, into three periods: (1) polemics during the life of Hegel (1816–31), (2) controversies in the religious field (1831–39), and (3) political debates (1840–44), though discussions on all of the problems continued through all three periods.

Polemics during the life of Hegel: 1816–31. While Hegel was still living, discussion was dominated by the master. It was not a matter of polemics within the school but only one of objections against the system from various quarters: from speculative theists; from Johann Herbart, a prominent student of the philosophy of mind, and his followers; and from disciples of Friedrich Schelling, an objective and aesthetic Idealist, and of Friedrich Schleiermacher, a seminal thinker of modern theology.

The substantive history of the school stems from Hegel's later teaching at Berlin and from the publication of his *Naturrecht und Staatswissenschaft im Grundrisse* (1821; Eng. trans., *The Philosophy of Right,* 1942). This book was reviewed by Herbart, who reprimanded Hegel for mixing the monism of the Rationalist Spinoza with the transcendentalism of Kant, which had explored the conditions of the possibility of knowledge in general. There were also certain critics who directed the liberal press against Hegel for attacking Jakob Fries, a psychologizing Neo-Kantian, in the introduction of *The Philosophy of Right.* Some of the polemical writings of Hegel made a notable impact—*e.g.,* a preface that he wrote for a book by one of his earliest disciples, Hermann Hinrichs, on the relation of faith to reason (1822). In this preface, Hegel saw the two things

as the same in content but different in form—which for faith is the representation and for reason is the concept.

Articles in *Jahrbücher*

Particularly significant were eight articles in the *Jahrbücher für wissenschaftliche Kritik* (founded 1827; "Yearbooks for Scientific Critique"), a journal of the Hegelian right. Important among these were a review by Hegel that was unexpectedly eulogistic about the thesis that philosophy and evangelical orthodoxy are compatible and another review in which Hegel responded indirectly to arguments of Herbart. Among Hegel's critics can be distinguished speculative theists such as Christian Weisse of Leipzig and Immanuel Fichte, the son of the more famous Johann Fichte, who reproached him for his panlogism and proposed to unify thought and experience in the concept of a free God, the Creator. Among the most loyal disciples of Hegel were Hermann Hinrichs, his collaborator, and Karl Rosenkranz, who defended the Hegelian solution of the faith–reason problem (which had asserted the identity of content and difference of form), thus aptly defending the free rationality of religion.

Period of controversies chiefly in religion: 1831–39. The tone of these early polemics became animated and embittered after the death of Hegel. But, inasmuch as conditions in Germany, during the Restoration, inhibited the liberalization of political discussions, the milieu of controversy shifted to the religious realm and became related to problems of immortality, Christology, and general theology.

Feuerbach and Strauss

Shortly before Hegel's death, the youthful Ludwig Feuerbach, who later became a pioneer of naturalistic humanism, had published his *Gedanken über Tod und Unsterblichkeit* (1830; "Thoughts on Death and Immortality"), in which he contended that, from the Hegelian point of view, death must be necessary in order for man to be transformed from the finite to the infinite and it is thus a privilege for man preferable to empirical personal survival. This work was held to confirm the charge of pantheism that orthodox adversaries had directed at Hegel's system. On this point, at the appearance of two volumes by Johann Friedrich Richter, a pantheist and critic of religion, Hegel's disciples intervened, in an argument employing not a few dialectical artifices, to conciliate Hegelian statements with the traditional doctrine of immortality.

The polarization of historical positions that the debate on immortality could not adequately express soon came into the open with *Das Leben Jesu kritisch bearbeitet* (1835–36; Eng. trans., *The Life of Jesus Critically Examined,* 1846), of David Friedrich Strauss, a biblical interpreter and radical theologian. This work brought the problem of the nature of Christ up to date from the point of view that had been reached by biblical criticism; *i.e.,* Christology was no longer an issue of denominational dogma but, rather, a problem of the interpretation and evaluation of the Gospel sources and of their meaning in the historical development of civilization. In this approach, the narrowly philological outlook was overcome by a reconstruction in terms of a philosophy of history strangely suggestive of the young Hegel. The thesis of the book was that the Gospel account is interwoven with myths that are not the works of individuals but of the collective poetic activity of the first Christian community, myths that resulted in part from messianic expectations, in part from the memory of the historical figure of Jesus, and in part from a transfiguration of the real elements. The aim of the myths was to demonstrate that philosophy and religion are the same in content and to offer, in an imaginative guise (as in parables), the meaning of the one truth that Substance is unification of the divine nature and of the human, which Christ symbolized and which is realized in the spirit of all humanity.

Strauss's work provoked a lively reaction, to which he replied in his *Streitschriften* (1837–38; "Controversial Writings"), proposing the image of a Hegelian school split, like the French Parliament, into a right (Göschel, and several others), a centre (Rosenkranz), and a left (Strauss himself). There were responses from the right and centre and from Bruno Bauer, a philosopher, historian, and biblical critic. From the anti-Hegelian side there was, above all, *Die evangelische Geschichte* (1838; "The History of the Gospels"), by Weisse, who, conceding to Strauss the necessity to rationalize the Gospel story, propounded a speculative interpretation of the Christ figure as an incarnation of the Logos (Thought-Word), in contrast to the mystic and pantheistic views.

Weisse, Ruge, Schelling

Meanwhile, Bauer shifted toward the left in a polemic against the orthodox Ernst Hengstenberg, a vehement accuser of the Hegelians, and in his *Kritik der Geschichte der Offenbarung* (1838; "Critique of the History of Revelation"). In 1838 was founded the earliest journal of the left, the *Hallische Jahrbücher für deutsche Wissenschaft und Kunst* ("Halle Yearbooks for German Science and Art"), coedited by the activist philosopher Arnold Ruge and T. Echtermeyer. At first, the journal maintained a moderate tone, and Hegelians of the centre and right also contributed articles. In June, however, it veered to the democratic-liberal side as Ruge struck out against an accuser of the young Hegelians and as Feuerbach attacked earlier Hegelians. Hegelianism, which marks the culmination of speculative philosophy, Feurebach charged, does not demonstrate its own truth, because its contrast between sensory reality and intellectual concept comprises an irresoluble contradiction. Thus, its dialectic turns out to be a "monologue with itself," bereft of authentic mediation with the world. Hegelian philosophy, he held, is a "rational mystique," and what is needed is a return to nature, which, as objective reason, ought to become a principle of philosophy and of art. Thus an extensive examination of contemporary culture was conducted by the journal's editors in an article that depicted Romanticism as a movement degraded to a reactionary stance and extolled the spirit of reform and of liberal (yet loyalist) Prussianism.

As for issues in the fields of logic and metaphysics, after several polemical exchanges the interest of philosophers was attracted to the publicist reawakening that came to Schelling, who reactivated certain anti-Hegelian criticisms. These criticisms dealt with the impossibility of building a valid philosophy upon the pure concept assumed as a point of departure and endowed with autonomous movement. Such a philosophy would be vitiated by presuppositions of what ought to be demonstrated and by hypostatizations (*i.e.,* the making of an idea into an entity). Schelling proposed, on the other hand, that the real itself be taken as the subject of development, to be grasped with a "lively intuition"; and that, while accepting a "negative philosophy" (such as that of Rationalism and Hegel) pointing to the conditions *without which* one cannot think, one must also add a "positive philosophy" delineating the conditions *by means of which* thought and reality can exist, premised on the existence of a free creative God.

Period of atheistic and political radicalism: 1840–44. The ensuing years marked one of the most intense periods in the cultural life of modern Europe.

Anti-Hegelian criticism. Advancing from Aristotelian presuppositions, an important critique against the Hegelian logic was presented by the classical philosopher and philologist Friedrich Adolf Trendelenburg in his *Logische Untersuchungen* (1840; "Logical Investigations"). In Hegel's view, the passage from Being to Nothing and to Becoming can be posited as a pure beginning "without presuppositions" of logic. In Trendelenburg's view, however, this passage is vitiated by its spurious dependence upon the surreptitious presupposition of the Empirical movement, without which support neither the passage from Being to Nothing (and vice versa) nor the recognition of Becoming as the "truth" of this primal opposition of concepts can be justified. Secondly, he charged that Hegel confused (1) the logical opposition or contradiction of *A* against non-*A* with (2) the real contradiction or contrariety of *A* against *B*. Contradiction (1) consists in the mere repetition of the first term with a negative sign; and from it no concrete movement can proceed. In contrariety (2), however, the opposition of the second term to the first is concrete—thus the second term cannot be deduced from the first and, instead, should be derived on its own account from empirical experience. Thus Hegel constructed his entire system, Trendelenburg charged, on an arbitrary dialectic of elements intrinsically real (contraries), which he mistakenly treated as though they were abstract opposites (contradictories) and were such by logical necessity.

Meanwhile, Schelling continued to teach his "positive philosophy"—of mythology and of revelation (of a personal God). Hence the philosophy of the later Schelling became the target of all of the criticisms from the left and likewise exerted a notable influence on the speculative theists. Meanwhile, the centre, on account of the critique of Trendelenburg, oriented itself toward the future reforms of Hegelianism.

Kierkegaard and Fischer

Among those who attended Schelling's lectures was Søren Kierkegaard, the man who was destined to become one of the founding fathers of Existentialism and whose religious individualism represents the earliest major result of the diffusion of Hegelianism outside of Germany. In all of his works—but above all in his *Philosophiske Smuler* (1844; Eng. trans., *Philosophical Fragments,* 1936) and his *Afsluttende uvidenskabelig Efterskrift* (1846; Eng. trans., *Concluding Unscientific Postscript,* 1941)—Kierkegaard waged a continuous polemic against the philosophy of Hegel. He regarded Hegel as motivated by the spirit of the harmonious dialectical conciliation of every opposition and as committed to imposing universal and panlogistic resolutions upon the authentic antinomies of life. Kierkegaard saw these antinomies as emerging from the condition of the individual, as a single person, who, finding himself always stretching to attain ascendance over his existential limitations in his absorption in God and at the same time always thrust back upon himself by the incommensurability of this relationship, cannot find his salvation except through the paradoxical inversion of the rational values of speculative philosophy and through the "leap of faith" in the crucified Christ. Kierkegaard's claim that the nexus of problems characterizing man's condition as an existing being is irreducible to any other terms lay at the very roots of Existentialism. It was destined to condition the critical relationship of this current of thought to Hegelianism throughout its subsequent history. Moreover, Kierkegaard's thought, which Kierkegaard did not know—still more than that of Strauss—seemed reminiscent of those problem areas explored in the young Hegel's religious thought—issues that were destined to appear only later when Hegel research would gain precise knowledge of the writings of Hegel's youth.

At this time the attitude of the centre was oriented toward reforms of the Hegelian system in the field of logic and historiography, as reflected especially in the emergence of Kuno Fischer, one of the foremost historians of philosophy. In the fundamental triad of the dialectic, as Fischer saw it, Being and Nothing are not equally static and neutralizing. The real movement does not interpose itself into their relationship because Being is here to be understood as the Being of thought, which, to the degree that it is a thinking of Nothing, possesses that dynamic surplus that becomes manifest in the moment of Becoming. It was in making responses to this view that the forthcoming Neo-Hegelian movement in Europe found some of its motivations.

Theological radicalism. In 1840 political conditions in Germany changed with the succession of the young Frederick William IV, whose minister began to repress the liberal press and summoned to Berlin in an anti-Hegelian capacity both Schelling and the conservative jurist F.J. Stahl, a stubborn critic of Hegel. Far from weakening the movement, however, these actions radicalized its revolutionary manifestations. Strauss, in *Die christliche Glaubenslehre* (1840–41; "The Christian Doctrine of Faith"), reaffirmed the opposition of philosophical pantheism to religious theism as a means of reunifying the finite and the infinite;

Feuerbach and Bauer

and Feuerbach established a philosophical anthropology in his major work *Das Wesen des Christentums* (1841; Eng. trans., *The Essence of Christianity,* new ed. 1957), in which man reappropriates his essence, which he had alienated from himself by hypostatizing it in the idea of God. The essence of man is reason, will, and love; and these three faculties comprise the consciousness of the human species as a knowledge of the infinity that man must regain. Man must thus reverse the theological propositions that express the spurious objectification of his universality in God; for this objectification had been effected through the individual consciousness in its effort to surmount its limitations. Thus Feuerbach interpreted the Christian mysteries as symbols of the alienation of human properties absolutized as divine attributes, and he criticized the contradictions of theology that are found in such concepts as God, the Trinity, the sacraments, and faith. Man's reappropriation of his essence from such religious alienation is consummated in the "new religion" of humanity, of which the supreme principle is that "man is God to man."

To this period belong also the major critiques of Bruno Bauer on the Johannine (1840) and Synoptic (1841–42) Gospels. Differentiating his position from the pantheistic and mysticizing Substance of Strauss, Bauer held that the Gospels were not the unconscious product of the original community but a product of the self-consciousness of the Spirit in a given stage of its development. There followed two works specifically concerning Hegel, in which, feigning an orthodoxy from which he charged Hegel with atheism and radicalism, Bauer maintained, in the form of a parody, the revolutionary interpretation of Hegel that became customary in the current of the Hegelian left.

Sociopolitical radicalism. In the years 1841–43, the repressive measures of the government reached ever more decisive extremes: Bauer was debarred from teaching; Feuerbach did not even attempt to teach; and Ruge was enjoined to publish the *Hallische* in Prussia instead of Leipzig. (Actually, he transferred it to Dresden and changed its name to the *Deutsche Jahrbücher.*) Here also appeared one of Ruge's major writings, "Die Hegelsche Rechtsphilosophie und die Politik unserer Zeit" (1842; "The Hegelian Philosophy of Right and the Politics of our Time"), in which Ruge denounced Hegel's political conservatism, charging that his contemplative reason was reduced to the acceptance of existing conditions, to the exclusion of every effort to modify reality, and to the absolutizing of the Prussian state as the model of an ideal state. Ruge's journal was suppressed early in 1843, but in March he published in Switzerland his *Anekdota zur neuesten deutschen Philosophie und Publicistik* ("Anecdotes for the Latest German Philosophy and Political Journalism"), containing articles by Bauer, Ruge, Marx, Feuerbach, and others.

Feuerbach's article developed the claim that the method of speculative philosophy, which is the ultimate form of theology, is to invert the subject and predicate—*i.e.,* to substantialize the abstract and to treat concrete determinations as attributes or "logical accidents" of hypostatized abstractions. The inversion of speculative propositions, he held, leads to the philosophical reappropriation of man's essence; the philosophy of the future will achieve mastery through the negation of the Hegelian philosophy—and this is exactly what he entitled his forthcoming book: *Grundsätze der Philosophie der Zukunft* (1843; "Basic Principles of the Philosophy of the Future"). In place of the immediate Absolute of Hegel, he argued, there must be substituted the immediate individual existent—corporeal, sensible, and rational. Man's reappropriation of himself will be possible whenever his need to transcend his own limitations finds fulfillment in another person and in the totality of the human species: "thus man is the measure of reason."

Meanwhile, a schism had been ripening in the left wing: (1) On the one hand, there were the "Free Berliners" (initially the young Friedrich Engels, later to become Marx's theoretician, the radical anarchist Max Stirner, and the Bauer brothers), who, deeming themselves faithful to Hegel, developed a philosophy of self-consciousness (understood in a subjective and superindividualistic sense) directed toward treating social and historical problems with aristocratic intellectual detachment. (2) On the other hand, there was the group that included Ruge, the publicist Moses Hess, the scholarly poet Heinrich Heine, and Karl Marx. Influenced in their theories by Feuerbach, this group directed radicalism toward an experience deepened by the classical Enlightenment and embraced the rising Socialism. They thus involved Hegel in their critique of the political, cultural, and philosophical conditions of the time. The most widely known result of the first trend was Stirner's book *Der Einzige und sein Eigentum* (1845; "The

Individual and His Property"), in which the fundamental thesis of individualistic anarchism can be discerned. The unique entity, in Stirner's view, is the individual, who must rebel against the attempt made by every authority and social organization to impose upon him a cause not his own and must be regarded as a focus of absolutely free initiative—a goal to be reached by emancipating himself from every idea-value imposed by tradition.

The work of Marx

The years between 1840 and 1844, however, saw the emergence of a figure incomparably more representative of the crisis of German Hegelianism than any already cited, that of Karl Marx, who was destined to guide the experience of this crisis toward a revolution of world historical scope. Marx's study of Hegel dates from his university years in Berlin, the earliest result of which was his doctoral dissertation with the exceedingly important preparatory notes, in which he ventured an original application of Hegelian method to the problem of the great crises in the history of philosophy. At first a friend of Bauer, Marx clung closely, however, to the democratic wing of the left. In 1843 he completed an important critical study of Hegel's *Philosophy of Right,* in which he reproached Hegel for having absolutized into an ideal state the Prussian state of the time. Such absolutizing, he charged, lent itself to generalizations of broad critical scope with respect to the idealistic procedure of hypostatizing the Idea and brought about (as allegorical derivatives from it) certain concrete political and social determinations, such as family, classes, and the state powers. Not yet a Communist, Marx nonetheless completed, in his *Kritik der hegelschen Staatsrechts* (written in the summer of 1843, published 1929; "Critique of Hegel's Constitutional Law"), a criticism of the erroneous relationship initiated in Hegel between society and the state, which was destined to lead Marx from the criticism of the modern state to that of modern society and its alienation.

It will be recalled that Hegel had likewise proposed the concept of alienation, describing the dialectic as a movement of the Absolute that was determined by its alienating and then regaining itself (thus overcoming the self-negation). Already in the *Economic and Philosophic Manuscripts of 1844* (German ed., 1932; Eng. trans., 1959), Marx had enunciated a general critique of the Hegelian dialectic that revealed its a priori nature, which, in Marx's view, was mystifying and alienated inasmuch as Hegel did nothing but sanction, by a method inverted with respect to real relationships, the alienation of all the concrete historical and human determinations.

Marx then directed himself against his former colleagues on the left—against Bauer in his *Die heilige Familie* (1845; Eng. trans., *The Holy Family,* 1956) and against Stirner in his *Die deutsche Ideologie* (1845–46; Eng. trans., *The German Ideology,* 1938), criticizing their "ideologism" (*i.e.,* the illusion that Idealism can be carried into the revolutionary camp since it is ideas that make history). The historical Materialism that Marx counterposed against Idealism expressed the conviction that the basis comprising the relations of production, both economic and social, conditions the superstructure of political, juridical, and cultural institutions and that the interchange among these spheres of production within the totality of an historical epoch must be designed to overcome their contradictions. This Materialism, though not belonging any more to Hegelianism, was destined nonetheless to remain linked to it by continuing polemical relationships and overlapping problem areas throughout the subsequent history of the movement.

Along with Marx must, of course, be mentioned his colleague Friedrich Engels, who was more tied, however, to the Hegelian conception of the dialectic—particularly regarding the dialectic of nature—than Marx was.

HEGELIANISM THROUGH THE 20TH CENTURY

Development and diffusion of Hegelianism in the later 19th century. In Germany, the second half of the 19th century witnessed a decline in the fortunes of Hegelianism, beginning with the *Hegel und seine Zeit* (1857; "Hegel and His Age"), by Rudolph Haym, a historian of the modern German spirit. The decline was urged on by Neo-Kantianism and Positivism as well as by the political realism of Bismarck. Hegelian influences still appeared in the first representatives of historicism (which urged that all things be viewed in the perspective of historical change). The surviving Hegelians, however, such as Kuno Fischer and Johann Erdmann, devoted themselves to the history of philosophy. Strauss and the Bauer brothers were won over to conservatism, and even Ruge, returning from exile in England, became a conservative.

Decline in Germany

Political and cultural problems: East Europe and the United States. The diffusion of Hegelianism outside of Germany was oriented in two directions. With respect to its political and cultural problems, the Hegelian experience developed in east European philosophers and critics such as the Polish count Augustus Cieszkowski, a religious thinker whose philosophy of action was initially influenced by the left; and the theistic metaphysician Bronislaw Trentowski. Among the Russians can be cited the literary critic Vissarion Belinsky, the democratic revolutionary writers Aleksandr Herzen and Nikolay Chernyshevsky, and certain anarchists such as the Russian exile and revolutionist Mikhail Bakunin. And among the French there were Hegelian Socialists such as Pierre-Joseph Proudhon.

In the United States, the interest in Hegelianism was stimulated by its political aspects and its philosophy of history. Its two centres, the St. Louis and Cincinnati schools, seemed to duplicate the German schism between a conservative and a revolutionary tendency. The former was represented by the Hegelians of the St. Louis school: the German Henry Brokmeyer and the New Englander William Harris, a pedagogue and politician, and the circle that they founded called the St. Louis Philosophical Society, which published an influential organ, *The Journal of Speculative Philosophy.* Their legitimism, or support for legitimate sovereignty, was expressed in the quest for a foundation, dialectical as well as speculative, for American democracy and in a dialectical interpretation of the history of the United States. The Cincinnati group, on the other hand, gathered around August Willich, a former Prussian officer, and John Bernard Stallo, an organizer of the Republican Party. Willich had participated in the Revolution of 1848 as a democratic partisan in south Germany, and, as an exile, had been in lively intercourse with Marx. He founded the *Cincinnati Republikaner,* in which he reviewed Marx's *Zur Kritik der politischen Ökonomie* (1859) and endeavoured to base the principles of social democracy upon the humanistic foundations of Feuerbach. Stallo, on the other hand, tried to interpret the political philosophy of Hegel in republican terms. The democratic community became, for him, the realization of the dialectic rationality of the Spirit with a rigorous separation of church and state.

St. Louis and Cincinnati schools

Logic and Metaphysics problems: Italy, England. The second trend in non-German Hegelianism was directed, in Italy and in England, to problems of logic and metaphysics. A vigorously speculative rethinking of the foundations of Hegel's *Wissenschaft der Logik* was engaged in by the major liberal Italian philosopher Bertrando Spaventa and his associates. Spaventa's *Studi sull' etica di Hegel* (1869) consisted of a direct liberal translation of Hegel's *Philosophy of Right.* Seeking to rediscover the connection between the thinking of the Italians of the 16th century and that of the German Idealists, Spaventa encountered the system of problems involved in the relationship between Kant and Hegel. He adopted from Kuno Fischer the solutions by which Fichte, Schelling, and Hegel had rendered Kant's transcendental ego consummatively veritable. He thus proposed an epistemological (idealistic theory of knowledge) interpretation of the Hegelian logic, according to which one premise of the logic is the dialectic of consciousness described in Hegel's *Phenomenology of Mind,* and the problems of the genesis of logic are resolved in the sense that Being is, from first to last, Becoming; *i.e.,* it is thought in action, which negates the objective residue of thought-out Being and, for that reason, is confirmed as a creative process. From Spaventa, whose intention was to vindicate the freedom and autonomy of thought against denominational dogmatism, was derived the foundation for the subjectivistic formalization

of Hegelianism soon undertaken by Giovanni Gentile, an early-20th-century Idealist.

As in Italy, so also in England, interest in Hegel arose from the philoopher's need to round out his experience of classical German thought by tracing its vicissitudes since the time of Kant; and this interest was directed toward the fields of epistemology and logic and in this instance was applied to problems of religion and not of politics. The pioneer in English Hegelianism was James Hutchison Stirling, through his work *The Secret of Hegel* (1865). Stirling reaffirmed the lineage of thought that Fischer had traced "from Kant to Hegel," endeavouring to penetrate the dialectic-speculative relationship of unity in multiplicity as the central point of the dialectic. Toward Hegelianism as a unifying experience the ethics scholar Thomas Hill Green, the foremost representative of Hegelianism at Oxford, applied himself, though with more original attitudes; and the brothers John and Edward Caird dedicated themselves to right-wing interpretations of religious subjects—Edward in a well-known monograph entitled *Hegel* (1883).

Hegelianism in the first half of the 20th century. At this point, the development of Hegelianism branched out in two directions: one of which, in England and Italy, pursued the tendencies of the Neo-Hegelians of the preceding decades, while the other, in Germany and France, accomplished the philological interpretative renewal known as the Hegel renaissance.

Bradley, Royce, Croce

Neo-Hegelianism in England and Italy. With respect to the first tendency, there appeared in England at the turn of the century various outstanding works on Hegel's logic by authors who were partly Hegelian in spirit. These scholars, toiling through the system of problems that they shared—which focussed on establishing a criterion for the unification of the multiplicity of experience—ended up in diverse positions: those of Bernard Bosanquet and John Ellis MacTaggart, for example, who were translators and commentators of Hegelian works; but above all that of the foremost spiritualistic philosopher then in England, F.H. Bradley, author of the renowned *Appearance and Reality* (1893), whose development led him to positions more and more at odds with the absolute panlogism of Hegel. His affirmation of the dualism of appearance and reality was the result of a critique of the category of relations, which, by introducing contradictions between the qualities of the thing, utterly shattered the unity of experience in which it might seem that true reality could be reached—a reality that in Bradley's view it is not given to thought to attain.

The echoes of this Idealistic system were not long in being felt in the United States by one of its most profound philosophers, an absolute Idealist, Josiah Royce, who, in *The World and the Individual* (1900–01), discussed the skeptical Idealism of Bradley in order to overthrow its consequences in favour of a conception of the infinite as a self-representative system and of the world (or the All) as an individualized realization of the intentional aims of the Idea copresent in a superior eternal consciousness. In Anglo-Saxon Neo-Hegelianism, the Hegelian experience has always been merely an episode—which fact serves to refine, by contrast, the methods of experimentalism that are more congenial to the Empirical tradition in England.

In Italy, on the other hand, the Neo-Hegelianism of the 20th century took the form of a spiritualistic reaction to the spread of Positivism that had followed upon the unification of Italy. This reaction developed in two directions: that of the historicism of Benedetto Croce and that of the actualism of Giovanni Gentile, two scholars who divided the realm of philosophy between themselves and occupied it—rather heavy-handedly—for four decades. The Crocean reform of Hegelianism dates from his volume *Ciò che è vivo e ciò che è morto della filosofia di Hegel* (1907; "What Is Living and What Is Dead in the Philosophy of Hegel") and from the systematic works of his so-called "philosophy of the spirit." Croce accepted the dialectic from Hegel as a requirement for the unification of opposites; but he rejected its system, in which Hegel would put in opposition and treat dialectically certain intellectual forms that are not really opposite but only distinct—such as the beautiful, the true, the useful, and the good, each of which has its dialectical opposite over against itself that it has to overcome within the purview of each grade. Consequently, renouncing the possibility of a philosophy of nature or of history, Croce formulated a development of so-called "distinct grades" according to the spiritual forms of art, of philosophy, of economics, and of ethics and contended that the comprehensive meaning of the development of the Spirit is given by history "as thought and as action" and a realization of freedom.

Gentile, on the other hand, accentuated the opposition of subject and object by considering every objective factuality as surpassed by the living dialectical development of the act—*i.e.*, the becoming of the Spirit in its own self-making, proceeding from an originating self-establishment, or *autoktisis,* of the Spirit itself. From this position he derived an absolute subjectivism that exploited all the possibilities for dialectically transforming every fixed position into its opposite, a downright sophistry of disengagement. Gentile's pro-Fascist stance, however, condemned his actualism to collapse.

Hegelian renaissance in Germany and France. Already from the beginnings of the century, however, there had been in Germany a change in Hegelian interpretation instigated by Wilhelm Dilthey's re-examination, in 1905, of the youthful manuscripts of Hegel and by the publication by one of Dilthey's principal disciples, Herman Nohl, of *Hegels theologische Jugendschriften* (1907; "The Theological Writings of Hegel's Youth"). Inasmuch as there had been heretofore only fragmentary notices on these unpublished literary remains, the effect of this rereading of the texts was to place them in contrast with the works of his maturity; they thus emerged as dealing, for the most part, with various problem areas in ethics, religion, and history; as lacking systematic preoccupations; and as rich discourse, tending to the mystic, which invited their comparison with the severe technical uniformity of his major works. Hermeneutical interest, however, centred especially on the problem of the beginnings of the philosophy and dialectic of Hegel, of which the first formulations were investigated in order to collate their meanings with those of the major works and of the *Phenomenology,* which was a key work of the Hegelian evolution inasmuch as it participated both in the romanticized colouring of the youthful writings and in the systematic demands of the *Encyklopädie der philosophischen Wissenschaften im Grundrisse* (1817; "Encyclopaedia of the Philosophical Sciences in Outline").

Publication of the *Jungendschriften*

Scholars were soon led to investigate the historical matrices of Hegel's intellectual culture—the late Enlightenment and dawning Romanticism—a direction of inquiry that yielded imposing contributions rich in discussions that continue to this day. These studies began with Dilthey's monograph, which pointed out the irrationalistic and vitalistic aspects of Hegel's youthful writings. In addition, a basic work by Franz Rosenzweig, *Hegel und der Staat* (1920), genetically reconstructed the political thought of the young Hegel in relation to its historical sources and concluded that the influence of Rousseau prevented Hegel from becoming the genuine "national philosopher of Germany." Jean Wahl, a French metaphysician and historian of philosophy, wrote on the "wretched conscience," interpreting Hegel existentially. Further, the German philosopher Richard Kroner studied the development from Kant to Hegel integrating it with the contributions of early Romanticism. And Hermann Glockner, a Bavarian aesthetic intuitionist, saw following one another in the development of Hegel a so-called "pantragistic" phase up to the *Phenomenology* and, subsequently, an opposing "panlogistic" phase that betrayed the most lively and concrete instances of the preceding phase—a work that approached the efforts at interpreting Hegel that were made by the Nazis.

Hegelian studies today. Today one has to speak not of the presence of Hegelianism as an operating philosophical current but only of studies on Hegel and of an experience of the Hegelian philosophy, to which, however, almost none of the present-day orientations in philosophy is foreign. The repeated encounter of Western culture with Marxist thought after World War II has brought to the fore the political, ethical, and religious implications of Hegelianism; and a marshalling into opposing camps analogous to that of the earlier crisis of the school is

taking shape. Today there are no orthodox Hegelians, but there are denominational critics of Hegelianism, especially Catholic, whose cognizance of Hegel's painful development invokes, despite their differences, a certain fellow feeling with him.

In the centre are found scholars of a liberal and radical frame of mind but with varying orientations with respect to historical interpretations. Karl Löwith, a German philosopher of history and culture, sees Hegel as the initiator of the "historicist" crisis in modern thought, culminating in Marx and in Kierkegaard; and to this he contrasts the metahistorical perspective reflected in the Nietzschean motif of the "eternal return," based on the ideal of a Goethean serenity. In France, Alexandre Kojève, noteworthy for his effort to harmonize Hegel with Martin Heidegger, proposes a reinterpretation of the *Phänomenologie* as a manifesto of the emancipation of "man the servant" from all alienations. Jean Hyppolite, author of an outstanding commentary on the *Phänomenologie,* usually presents a restrained humanistic interpretation of the Hegel of Jena. This renaissance of the study of Hegel has conditioned the thought of some of the major thinkers of France. Particularly notable, however, is the Hegelian conditioning of German philosopher-sociologists such as Theodor Adorno and Herbert Marcuse. The former is sometimes regarded as the most Hegelian thinker of the mid-20th century because he sought to bring again to the fore Hegel's dialectic, understood in a new anti-intellectualistic sense, as a method for the solution of present-day social problems. Marcuse, a partisan of a Diltheian interpretation, approaches the position of the first Hegelian left, ending up in what critics see as a neoromantic anarchism. The major merit of both of these thinkers lies in their incisive analyses of aspects of modern consumer societies, especially American—though their proposed remedies remain uncertain.

Three currents of Marxist interpretation

The major interest, however, in the contemporary interpretation of Hegel is displayed by the Marxist camp. Marxist interpretation of Hegel had permeated the entire history of Hegelianism (notwithstanding the fact that the critical activity of young Marx against Hegel had been vehemently conducted and had led to various effects). This interpretation had settled upon the distinction made by Friedrich Engels between the method and the system of Hegel's philosophy—*i.e.,* between the dialectic considered as a revolutionary "principle of movement" that achieves fulfillment in human culture, and the system, regarded, on the other hand, as reactionary because idealistic and conservative. With varying emphases on critical issues, this interpretation was continued in subsequent Marxist thinkers—from the Russians Georgy Plekhanov and Lenin to Mao Tse-tung and Joseph Stalin—the latter of whom affirmed the complementariness of historical and dialectical Materialism.

Today many Marxist scholars, especially in the countries of eastern Europe, remain favourable to the traditional line of Engels; and above all György Lukács, a Hungarian philosopher and literary critic and author of a volume on the young Hegel, does so. With the intention of revealing the romantic and irrationalistic presuppositions of Naziism, Lukács reevaluates, in German culture, the tendency of the Enlightenment and of democracy, which he recognizes in the young Goethe, in Schiller, in Hölderlin, and in the young Hegel—in whom he sees, however, a reactionary involution.

A secondary tendency, which is drawing attention in France, with the work of Louis Althusser, draws Marx close to Structuralism, a recent school that seeks through a "human science," to probe the systematic structures evinced in cultural life. In this school Marx's humanism is viewed as a temporary, Feuerbachian phase, surpassed by commitment to the scientific observation of the structure of bourgeois society. Such Structuralistic interpretation of Marxism thus runs the risk of departing from a due emphasis on the *historical* substance of Marxian Materialism.

The latter motive is, on the other hand, the essential aim of a third Marxist current, in Italy, initiated by Galvano della Volpe, a critical aesthetician who discusses the relationship between bourgeois and Socialist democracy and champions, in aesthetics, a critical and antiromantic Aristotelianism. This current has been continued by Mario Rossi, who asks one to read again in full the texts of Hegel and Marx, to reconstruct the related movements, and to compare the Materialistic conception of history with more recent philosophical currents such as Structuralism, present-day sociology, and the logic of the sciences.

A conclusion of a theoretical-systematic nature concerning Hegelianism has today become not only impossible but also inopportune, because its possible interest has been effectively replaced by that of the sheer history of the movement. The latter has shown how the substantial ambiguity of the philosophy and dialectic of Hegel can be resolved only when its claim to be able to solve all problems on a theoretical level and to achieve a "circular" decisiveness in its arguments—which violates the conditioning specificity of historical facts—is refuted. It is then the scholar's task to explore the limits of Hegel's thought as well as its conditioned inadequacies—but also its merits, which are above all those of having expressed and documented the major part of the cultural problems of modern civilization. (M.R.)

BIBLIOGRAPHY

Hegel. *Works:* A collected edition of Hegel's published works, together with a great deal of material culled from his lectures, was published by his pupils within a few years of his death in 1831. This edition, with some rearrangement, was reissued by HERMANN GLOCKNER in 26 volumes, including a comprehensive index (1927–40). In 1905 the Philosophische Bibliothek (Leipzig, later Hamburg) began publication of a new edition with a carefully revised text edited by GEORG LASSON and later by JOHANNES HOFFMEISTER; volumes appeared for more than 50 years, but it was not completed. It has been enhanced by a comprehensive edition sponsored by the Deutsche Forschungsgemeinschaft, which is to contain about 50 volumes. The first volume appeared in 1968. English translations of most of Hegel's works were published in the late 19th and early 20th century, but, apart from those by WILLIAM WALLACE (*Logic* and *Mind—i.e.,* the first and third parts of the *Encyclopaedia*), they are not always satisfactory and they have no notes. With a view to remedying this deficiency, new English translations have appeared of some works, including *Philosophy of Right* (1942, often reprinted) and *Science of Logic* (1969), as well as translations of writings not translated previously, such as *Early Theological Writings* (1948; rev. ed., 1971), and *Philosophy of Nature,* 3 vol. (1970), the second part of the *Encyclopaedia.* With the exception of *Science of Logic* and the Oxford translation of *Philosophy of Nature,* all these translations are annotated.

Life and philosophy: RAYMOND PLANT, *Hegel* (1973), is a study of origins of his thought; and CHARLES TAYLOR, *Hegel* (1975), is a study of the development of his philosophy. An excellent short account of Hegel's philosophy in English is EDWARD CAIRD, *Hegel* (1883, reissued 1972), but it has been updated in certain respects by G.R.G. MURE, *The Philosophy of Hegel* (1965); and in more detail by WALTER A. KAUFMANN, *Hegel* (1965, reissued 1978). An attempt to interest modern philosophers in Hegel is contained in J.N. FINDLAY, *Hegel* (1958, reissued 1976), but this important and lively work is for consideration only by those already acquainted with Hegel. As an introduction, G.R.G. MURE, *An Introduction to Hegel* (1940, reprinted 1982), is more reliable but it is not an exposition. A standard long exposition of Hegel's mature system is KUNO FISCHER, *Hegels Leben, Werke und Lehre,* especially the 2nd ed. (1911, reprinted 1976); while in English there is WALTER T. STACE, *The Philosophy of Hegel* (1924). In 1900 Wilhelm Dilthey maintained that Hegel could be understood only if there were a study of his early manuscripts; on the basis of these, Dilthey wrote *Die Jugendgeschichte Hegels* (1906, reissued 1968), a history of Hegel's development. This seminal work, hardly noticed at all by writers in English before 1965, gave rise to an immense literature in Germany, France, and Italy. An important and brilliant study of the young Hegel is GEORG LUKACS, *Der junge Hegel* (1948; Eng. trans. 1975), written from a Marxist point of view. An exhaustive study is that of THEODOR HAERING, *Hegel: sein Wollen und sein Werk,* 2 vol. (1929–38, reissued 1979). Reliable and more readable than this are the two volumes of HERMANN GLOCKNER issued (1929 and 1940) as an appendix to his edition of the collected works. Even these, however, have been outdated by the flood of material collected by the Hegel-Archiv at Bochum in West Germany, and published in a series of volumes of *Hegel-Studien* (1961 and subsequent years), and by the four volumes of Hegel's letters, edited by JOHANNES HOFFMEISTER (1952–60). An admirable, but now obsolete, biography is by KARL ROSENKRANZ (1824).

Specialized commentaries: (*Mind*): JUDITH N. SHKLAR, *Freedom and Independence: A Study of the Political Ideas of Hegel's "Phenomenology of Mind"* (1976), is a guidebook and commentary. (*Phenomenology of Spirit*): JEAN HYPPOLITE, *Genèse et structure de la Phénoménologie de l'Esprit de Hegel* (1946, reissued 1970). (*Logic*): G.R.G. MURE, *A Study of Hegel's Logic* (1950, reissued 1967); STANLEY ROSEN, *G.W.F. Hegel: An Introduction to the Science of Wisdom* (1974), is a study of the development and meaning of his dialectic; and HANS-GEORG GADAMER, *Hegel's Dialectic: Five Hermeneutical Studies* (1976), is an analysis for specialists. (*Nature*): The apparatus in the English translation by M.J. PETRY, 3 vol. (1970), provides a full and learned commentary. (*Law, morality, and the state*): WILHELM SEEBERGER, *Hegel; oder, die Entwicklung des Geistes zur Freiheit* (1961), is a good introduction to Hegel's thought as a whole; but HUGH A. REYBURN, *The Ethical Theory of Hegel* (1921, reissued 1970); and FRANZ ROSENZWEIG, *Hegel und der Staat,* 2 vol. (1920, reissued 1962), are excellent summaries, and the latter is a commentary as well. SHLOMO AVINERI, *Hegel's Theory of the Modern State* (1972), is an account of his political thought. See also GEORGE A. KELLY, *Hegel's Retreat from Eleusis: Studies in Political Thought* (1978). (*Art*): JACK KAMINSKY, *Hegel on Art* (1962, reissued 1970), is a fair summary of Hegel's lectures. (*Religion*): THOMAS M. KNOX, *A Layman's Quest* (1969), deals, in chapters 5 and 6, not only with Hegel's lectures on the philosophy of religion but also with all his other writings elsewhere on religion; BERNARD M.G. REARDON, *Hegel's Philosophy of Religion* (1977), is a summary. (*History*): BURLEIGH T. WILKINS, *Hegel's Philosophy of History* (1974), is an introduction; and GEORGE D. O'BRIEN, *Hegel on Reason and History: A Contemporary Reinterpretation* (1975), questions his reputation as an anti-empirical, apriorist thinker.

Hegelianism. *Critical works:* Works presenting a critical consideration of Hegelianism viewed as a whole are few. See, however: STEPHAN D. CRITES, "Hegelianism," in the *Encyclopedia of Philosophy,* vol. 3, pp. 451–459 (1967, reissued 1972); MARIO ROSSI, *Da Hegel a Marx,* 2 vol. (1970); and RENÉ SERREAU, *Hegel et l'hégélianisme,* 4th ed. (1971).

Historical works: JOHN E. TOEWS, *Hegelianism: The Path Toward Dialectical Humanism, 1805–1841* (1980); JOHANN E. ERDMANN, *Die Deutsche Philosophie seit Hegels Tode* (1963); WILLY MOOG, *Hegel und die Hegelsche Schule* (1930, reissued 1973); KARL LÖWITH, *From Hegel to Nietzsche: The Revolution in Nineteenth-Century Thought,* 1964; originally published in German, 1941); and two anthologies, on the Left and Right, respectively: KARL LÖWITH (ed.), *Die Hegelsche Linke* (1962); and HERMANN LÜBBE (ed.), *Die Hegelsche Rechte* (1962).

In various countries: (*Germany*): HEINRICH LEVY, *Die Hegel-Renaissance in der deutschen Philosophie* (1927). *(Italy):* MARIO ROSSI (ed.), *Sviluppi dello Hegelismo in Italia* (1957); BENEDETTO CROCE, *Saggio sullo Hegel,* 5th ed. (1967). (*Slavic countries*): Contributions of authors from Russia, Poland, the Balkans, and Czechoslovakia are presented in *Hegel bei den Slaven,* 2nd ed., ed. by DMITRIJ TSCHIŽEWSKIJ (1961); see also BORIS JAKOWENKO, *Ein Beitrag zur Geschichte des Hegelianismus in Russland* (1934). (*England*): HIRA-LAL HALDAR, *Neo-Hegelianism* (1927). (*United States*): LOYD D. EASTON, "Hegelianism in Nineteenth-Century Ohio," *Journal of the History of Ideas,* 23:355-378 (1962), for the Cincinnati school; HENRY A. POCHMANN, *German Culture in America: Philosophical and Literary Influences 1600–1900,* pp. 257–294 (1957, reprinted 1978), for the St. Louis school. See also WILLIAM H. GOETZMANN and DICKSON PRATT (eds.), *The American Hegelians: An Intellectual Episode in the History of Western America* (1973).

Other works: AUGUSTE CORNU, *Karl Marx et Friedrich Engels,* 2 vol. (1955–58), is rich in materials and citations from the *Hallische* and *Deutsche Jahrbücher.* See also HERBERT MARCUSE, *Reason and Revolution: Hegel and the Rise of Social Theory,* 2nd ed. (1954); and SIDNEY HOOK, *From Hegel to Marx* (1936, reissued 1950 and 1962).

(T.M.K./M.R./Ed.)

Heisenberg

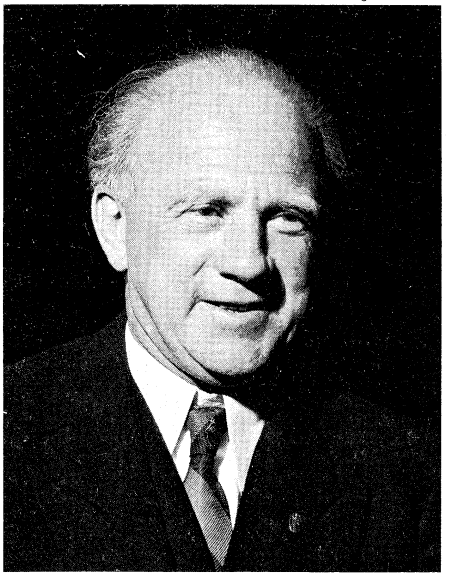
Werner Karl Heisenberg, physicist, philosopher, and public figure, helped to establish the modern science of quantum mechanics, out of which came the famous indeterminacy principle. He also made important contributions to the theories of the hydrodynamics of turbulence, the atomic nucleus, ferromagnetism, cosmic rays, and elementary particles; and he planned the first post-World War II German nuclear reactor, at Karlsruhe, West Germany.

Major influences

In his philosophical and methodological writings, Heisenberg was much influenced by Niels Bohr and Albert Einstein. From the former he derived the concepts of the social and dialogical character of scientific invention; the principle of correspondence (pragmatic and model-theoretical continuity) between macrophysics and microphysics; the permanence, though not the universality, of classical physics; the "interactive," rather than passive, role of the scientific observer in microphysics; and, consequently, the contextualized character of microphysical theories. From Einstein he derived the concepts of simplicity as a criterion of the central order of nature; scientific realism (*i.e.*, science describing nature itself, not merely how nature can be manipulated); and the theory-ladenness of scientific observations. He was coauthor with Bohr of the philosophy of complementarity. In his later work he conceived of a central order in nature, consisting of a set of universal symmetries expressible in a single mathematical equation for all systems of particulate matter. As a public figure, he actively promoted the peaceful use of nuclear energy after World War II and, in 1957, led other German scientists in opposing a move to equip the West German Army with nuclear weapons. He was, in 1954, one of the organizers of the Conseil Européen pour la Recherche Nucléaire (CERN; later, Organisation Européene pour la Recherche Nucléaire) in Geneva.

Early life. Heisenberg was born on December 5, 1901, in Würzburg, Germany. He studied physics, together with Wolfgang Pauli, his lifelong friend and collaborator, under Arnold Sommerfeld at the University of Munich and completed his doctoral dissertation (1923) on turbulence in fluid streams. Heisenberg followed Pauli to the University of Göttingen and studied there under Max Born; then, in the fall of 1924, he went to the Institute for Theoretical Physics in Copenhagen to study under Bohr.

Edo Koenig—Black Star



Heisenberg.

Heisenberg's interest in Bohr's model of the planetary atom and his comprehension of its limitations led him to seek a theoretical basis for a new model. Bohr's concept—after 1913 the centrepiece of what has come to be called the old quantum theory—had been based on the classical motion of electrons in well-defined orbits around the nucleus, and the quantum restrictions had been imposed arbitrarily to bring the consequences of the model into conformity with experimental results. As a summary of existing knowledge and as a stimulus to further research, the Bohr atom had succeeded admirably, but the results of new research were becoming more and more difficult to reconcile with it.

In June 1925, while recuperating from an attack of hay fever on Helgoland, an island in the North Sea, Heisenberg solved a major physical problem—how to account for the stationary (discrete) energy states of an anharmonic oscillator. His solution, because it was analogous to that of a simple planetary atom, launched the program for the development of the quantum mechanics of atomic systems. (Quantum mechanics is the science that accounts for discrete energy states—as in the light of atomic spectra—and other forms of quantized energy, and for the phenomenon of stability exhibited by atomic systems.) Heisenberg published his results some months later in the *Zeitschrift für Physik* under the title "Über quantentheoretische Umdeutung kinematischer und mechanischer Beziehungen" ("About the Quantum-Theoretical Reinterpretation of Kinetic and Mechanical Relationships"). In this article he proposed a reinterpretation of the basic concepts of mechanics.

Heisenberg's treatment of the problem departed from Bohr's as much as Bohr's had from 19th-century tenets. Heisenberg was willing to sacrifice the idea of discrete particles moving in prescribed paths (neither particles nor paths could be observed) in exchange for a theory that would deal directly with experimental facts and lead to the quantum conditions as consequences of the theory rather than ad hoc stipulations. Physical variables were to be represented by arrays of numbers; under the influence of Einstein's paper on relativity (1905), he took the variables to represent not hidden, inaccessible structures but "observable" (*i.e.*, measurable) quantities. Born saw that the arrays obeyed the rules of matrix algebra; he, Pascual Jordan, and Heisenberg were able to express the new theory in terms of this branch of mathematics, and the new quantum theory became matrix mechanics. Each (usually infinite-dimensional) matrix of the theory specified the set of possible values for a physical variable, and the individual terms of a matrix were taken to generate probabilities of occurrences of states and transitions among states. Heisenberg used the new matrix mechanics to interpret the dual spectrum of the helium atom (that is, the superposed spectra of its two forms, in which the spins of the two electrons are either parallel or antiparallel), and with it he predicted that the hydrogen molecule should have analogous dual forms. With others, he also addressed many atomic and molecular spectra, ferromagnetic phenomena, and electromagnetic behaviour. Important alternative forms of the new quantum theory were proposed in 1926 by Erwin Schrödinger (wave mechanics) and P.A.M. Dirac (transformation theory).

Indeterminacy principle

In 1927 Heisenberg published the indeterminacy, or uncertainty, principle. The form he derived appeared in a paper that tried to show how matrix mechanics could be interpreted in terms of the intuitively familiar concepts of classical physics. If q is the position coordinate of an electron (in some specified state), and p its momentum, assuming that q, and independently, p have been measured for many electrons (all in the particular state), then, Heisenberg proved,

$$\Delta q \cdot \Delta p > h,$$

where Δq is the standard deviation of measurements of q, Δp is the standard deviation of measurements of p, and h is Planck's constant (6.626176×10^{-27} erg-second). Indeterminacy principles are characteristic of quantum physics; they state the theoretical limitations imposed upon any pair of noncommuting (*i.e.*, conjugate) variables, such as the matrix representations of position and momentum; in such cases, the measurement of one affects the measurement of the other. The enormous significance of the indeterminacy principle is recognized by all scientists; but how it is to be understood physically—whether it depends on using intuitive classical ("complementary") pictures of a quantum system, or whether it is a principle in (a new kind of quantum) statistics, or whether in some sense through the special properties of the mathematical model it also describes a character of individual quantum systems—has been and still is much disputed. Bohr took the principle to apply to the complementary pictures of a quantum system—as a particle or as a wave pocket in classically intuited space; Heisenberg originally took the principle to apply to the nonintuitive properties of quantum, as distinct from classical, systems.

Bohr and Heisenberg elaborated a philosophy of complementarity to take into account the new physical variables and an appropriate measurement process on which each depends. This new conception of the measurement process in physics emphasized the active role of the scientist, who, in making measurements, interacted with the observed object and thus caused it to be revealed not as it is in itself but as a function of measurement. Many physicists, including Albert Einstein, Erwin Schrödinger, and Louis de Broglie, refused to accept the philosophy of complementarity.

Later life. From 1927 to 1941 Heisenberg was professor at the University of Leipzig. For the following four years, he was director of the Kaiser Wilhelm Institute for Physics in Berlin. Although he did not publicly oppose the Nazi regime, he was hostile to its policies. During World War II he worked with Otto Hahn, one of the discoverers of nuclear fission, on the development of a nuclear reactor. He failed to develop an effective program for nuclear weapons, probably from want of technical resources and lack of will to do so. After the war he organized and became director of the Max Planck Institute for Physics and Astrophysics at Göttingen, moving with the institute, in 1958, to Munich; he was also, in 1954, the German representative for the organizing of CERN.

Universal nonlinear spinor equation

In the postwar period Heisenberg began working on a fundamental spinor equation (a nonlinear differential equation capable of representing with spinors—complex vector-like entities—all possible particulate states of matter). His intuitions had led him to postulate that such an equation would exhibit a basic set of universal symmetries in nature (a symmetry is a mathematic form invariant under groups of canonical space–time and other changes in the representing elements), and be capable of explaining the variety of elementary particles generated in high-energy collisions. In this work, the "Platonic" character of which he recognized, he had the support and collaboration of Hans-Peter Dürr and Carl Friedrich von Weizsäcker.

Although he early, and indirectly, came under the influence of Ernst Mach, Heisenberg, in his philosophical writings about quantum mechanics, vigorously opposed the Logical Positivism developed by philosophers of science of the Vienna Circle. According to Heisenberg, what was revealed by active observation was not an absolute datum, but a theory-laden datum; *i.e.*, relativized by theory and contextualized by observational situations. He took classical mechanics and electromagnetics, which articulated the objective motions of bodies in space–time, to be permanently valid, though not applicable to quantum mechanical systems; he took causality to apply in general not to individual quantum mechanical systems but to mathematical representations alone, since particle behaviour could be predicted only on the basis of probability.

Heisenberg married Elisabeth Schumacher in 1937; they had seven children. He loved music in addition to physics and saw a deep affinity between these two interests. He also wrote philosophical works, believing that new insights into the ancient problems of Part and Whole and One and Many would help discovery in microphysics. Widely acknowledged as one of the seminal thinkers of the 20th century, Heisenberg was awarded the Nobel Prize for Physics for 1932 and was honoured with the Max Planck Medal, the Matteucci Medal, and the Barnard College Medal of Columbia University. He died in Munich on February 1, 1976. (P.A.H.)

BIBLIOGRAPHY. Books by Heisenberg include *Physical Principles of the Quantum Theory* (1930, reprinted 1950; originally published in German, 1930), his most important work, containing themes of early papers amplified into a treatise; *Philosophic Problems of Nuclear Science* (1952, reprinted 1966; originally published in German, 1942, enlarged ed. 1948), a collection of his early essays; *Physics and Philosophy: The Revolution in Modern Science* (1958, reprinted 1962), his Gifford lectures; *Physics and Beyond* (1971, reprinted 1972; originally published in German, 1969), a memoir of his early life; and *Across the Frontiers* (1974; originally published in German, 1971), collected essays and occasional lectures. Biographical material is found in ARMIN HERMANN, *Werner Heisenberg, 1901–1976* (1976); CARL F. VON WEIZSÄCKER, *Werner Heisenberg* (1977); ELISABETH HEISENBERG, *Das politische Leben eines Unpolitischen: Erinnerungen an Werner Heisenberg* (1980), and the works cited there. For Heisenberg's role in the German wartime atomic program, see LESLIE R. GROVES, *Now It Can Be Told: Story of the Manhattan Project* (1962, reprinted 1983). Collections in honour of Heisenberg include FRITZ BOPP (ed.), *Werner Heisenberg und die Physik unserer Zeit* (1961); HEINRICH PFEIFFER (ed.), *Denken und Umdenken: Zu Werk und Wirkung von Werner Heisenberg* (1977); and PETER BREITENLOHNER and H. PETER DÜRR (eds.), *Unified Theories of Elementary Particles: Proceedings of the Heisenberg Symposium, July 16–21, 1981* (1982). For the history of quantum mechanics, see THOMAS KUHN *et al.* (eds.), *Sources for History of Quantum Physics* (1967), interviews with the principal originators of the quantum theory, the transcripts and tapes of which are deposited in many places; and EDWARD M. MACKINNON, *Scientific Explanation and Atomic Physics* (1982), a detailed historical account, with an excellent bibliography. For studies of Heisenberg's philosophy of science, see PATRICK A. HEELAN, *Quantum Mechanics and Objectivity* (1965); MAX JAMMER, *The Philosophy of Quantum Mechanics: The Interpretation of Quantum Mechanics in Historical Perspective* (1974), and *The Conceptual Development of Quantum Mechanics* (1966), which provide the most complete study of Heisenberg's contribution to quantum mechanics.

Helmholtz

One of the greatest scientists of the 19th century, Hermann Ludwig Ferdinand von Helmholtz made fundamental contributions to physiology, optics, electrodynamics, mathematics, and meteorology, but he is best known for his statement of the law of the conservation of energy. In addition, he brought to his laboratory research the ability to analyze the philosophical assumptions on which much of 19th-century science was based, and he did so with clarity and precision.

Early life. Helmholtz, the eldest of four children, was born at Potsdam, near Berlin, on August 31, 1821, and because of his delicate health was confined to home for his first seven years. His father was a teacher of philosophy and literature at the Potsdam Gymnasium, and his mother was descended from William Penn, the founder of Pennsylvania. From his mother came the calm and reserve that marked him all his life. From his father came a rich, but mixed, intellectual heritage. His father taught him the classical languages, as well as French, English, and Italian. He also introduced him to the philosophy of Immanuel Kant and Johann Gottlieb Fichte and to the approach to nature that flowed from their philosophical insights. This "Nature philosophy," in the hands of early 19th-century investigators, became a speculative science in which it was felt that scientific conclusions could be deduced from philosophical ideas, rather than from empirical data gathered from observations of the natural world. Much of Helmholtz' later work was devoted to refuting this point of view. His empiricism, however, was always deeply influenced by the aesthetic sensitivity passed on to him by his father, and music and painting played a large part in his science.

By courtesy of the Ruprecht-Karl-Universitat, Heidelberg, Germany



Helmholtz.

After graduating from the gymnasium, Helmholtz in 1838 entered the Friedrich Wilhelm Medical Institute in Berlin, where he received a free medical education on the condition that he serve eight years as an army doctor. At the institute he did research under the greatest German physiologist of the day, Johannes Müller. He attended physics lectures, worked his way through the standard textbooks of higher mathematics, and learned to play the piano with a skill that later helped him in his work on the sensation of tone.

On graduation from medical school in 1843, Helmholtz was assigned to a regiment at Potsdam. Because his army duties were few, he did experiments in a makeshift laboratory he set up in the barracks. At that time he also married Olga von Velten, daughter of a military surgeon.

Early success

Before long, Helmholtz' obvious scientific talents led to his release from military duties. In 1848 he was appointed assistant at the Anatomical Museum and lecturer at the Academy of Fine Arts in Berlin, moving the next year to Königsberg, in East Prussia (now Kaliningrad), to become assistant professor and director of the Physiological Institute. But Königsberg's harsh climate was injurious to his wife's health, and in 1855 he became professor of anatomy and physiology at the University of Bonn, moving in 1858 to Heidelberg. During these years his scientific interests progressed from physiology to physics. His growing scientific stature was further recognized in 1871 by the offer of the professorship of physics at the University of Berlin; in 1882, by his elevation to the nobility; and, in 1888, by his appointment as first director of the Physico-Technical Institute at Berlin, the post that he held for the rest of his life.

The variety of positions he held reflects his interests and competence but does not reflect the way in which his mind worked. He did not start out in medicine, move to physiology, then drift into mathematics and physics. Rather, he was able to coordinate the insights he had acquired from his experience in these disciplines and to apply them to every problem he examined. His greatest work, *Handbook of Physiological Optics* (1867), was characterized—like all of his scientific works—by a keen philosophical insight, molded by exact physiological investigations, and illustrated with mathematical precision and sound physical principles.

The general theme that runs through most, if not all, of Helmholtz' work may be traced to his rejection of Nature philosophy, and the violence of his rejection of this seductive view of the world may well indicate the early attraction it had for him. Nature philosophy derived from Kant, who in the 1780s had suggested that the concepts of time, space, and causation were not products of sense experience but mental attributes by which it was possible to perceive the world. Therefore, the mind did not merely record order in nature, as the Empiricists insisted; rather, the mind organized the world of perceptions so that, reflecting the divine reason, it could deduce the system of the world from a few basic principles. Helmholtz opposed this view by insisting that all knowledge came through the senses. Furthermore, all science could and should be reduced to the laws of classical mechanics, which, in his view, encompassed matter, force, and, later, energy, as the whole of reality.

Attack on Nature philosophy

Helmholtz' approach to nature was evident in the very first scientific researches he undertook while working for his doctorate in the laboratory of Müller. Like most biologists, Müller was a vitalist who was convinced that it would be impossible ever to reduce living processes to the ordinary mechanical laws of physics and chemistry. The organism as a whole, he insisted, was greater than the sum of its physiological parts. There must be some vital force that coordinated the physiological action of organs to produce the harmonious organic behaviour that characterized the living creature. Such a vital force was not susceptible to experimental investigation, and Müller therefore concluded that a truly experimental physiology was impossible.

In Müller's laboratory Helmholtz met a group of young men, among whom were Emil Heinrich Du Bois-Reymond, the founder of experimental neurophysiology, and Ernst Wilhelm von Brücke, who later became an expert on the operations of the human eye. Du Bois-Reymond expressed their opposition to Müller's views in a statement that fully expressed Helmholtz' own position. "Brücke and I," Du Bois-Reymond wrote, "we have sworn to each other to validate the basic truth that in an organism no other forces have any effect than the common physiochemical ones. . . ."

Opposition to Müller's vitalism

It was with this attitude that Helmholtz began his doctoral thesis in 1842 on the connection between nerve fibres and nerve cells. This soon led him to a broader

field of inquiry, namely, the source of animal heat. Recent publications in France had cast doubt upon the earlier confident assertion that all the heat produced in an animal body was the result of the heats of combination of the various chemical elements involved, particularly carbon, hydrogen, and oxygen. In 1842 Justus von Liebig attempted to reestablish the mechanical theory of animal heat in his book *Animal Chemistry; or, Organic Chemistry in Its Application to Physiology and Pathology.* Liebig tried to do this by experiments, whereas Helmholtz took a much more general path. Having mastered both physics and mathematics, Helmholtz could do what no other physiologist of the time could even attempt—subject the problem to a mathematical and physical analysis. He supposed that, if vital heat were not the sum of all the heats of the substances involved in chemical reactions within the organic body, there must be some other source of heat not subject to physical laws. This, of course, was precisely what the vitalists argued. But such a source, Helmholtz went on, would permit the creation of a perpetual motion machine if the heat could, somehow, be harnessed. Physics, however, had rejected the possibility of a perpetual motion machine as early as 1775, when the Paris Academy of Sciences had declared itself on the question. Hence, Helmholtz concluded, vital heat must be the product of mechanical forces within the organism. From there he went on to generalize his results to state that all heat was related to ordinary forces and, finally, to state that force itself could never be destroyed. His paper "On the Conservation of Force," which appeared in 1847, marked an epoch in both the history of physiology and the history of physics. For physiology, it provided a fundamental statement about organic nature that permitted physiologists henceforth to perform the same kind of material and energy balances as their colleagues in physics and chemistry. For the physical sciences, it provided one of the first, and certainly the clearest, statements of the principle of the conservation of energy.

In 1850 Helmholtz drove another nail into the coffin of vitalism. Müller had used the nerve impulse as an example of a vital function that would never be submitted to experimental measurement. Helmholtz found that this impulse was perfectly measurable and had the remarkably slow speed of some 90 feet (27 metres) per second. (This measurement was obtained by the invention of the myograph and illustrates Helmholtz' ability to create new instruments.) The slowness of the nerve impulse further supported those who insisted that it must involve the rearrangement of ponderable molecules, not the mysterious passage of a vital force.

Invention of the ophthalmoscope and ophthalmometer

Among Helmholtz' most valuable inventions were the ophthalmoscope and the ophthalmometer (1851). While doing work on the eye, and incidentally showing that it was a rather imperfect piece of workmanship not at all consonant with the vitalistic idea of the divine mind at work, Helmholtz discovered that he could focus the light reflected from the retina to produce a sharp image of the tissue. The ophthalmoscope remains one of the most important instruments of the physician, who can use it to examine retinal blood vessels, from which clues to high blood pressure and to arterial disease may be observed. The ophthalmometer permits the measurement of the accommodation of the eye to changing optical circumstances, allowing, among other things, the proper prescription of eyeglasses.

Helmholtz' researches on the eye were incorporated in his *Handbook of Physiological Optics,* the first volume of which appeared in 1856. In the second volume (1867), Helmholtz further investigated optical appearances and, more importantly, came to grips with a philosophical problem that was to occupy him for some years—Kant's insistence that such basic concepts as time and space were not learned by experience but were provided by the mind to make sense of what the mind perceived. The problem had been greatly complicated by Müller's statement of what he called the law of specific nerve energies. Müller discovered that sensory organs always "report" their own sense no matter how they are stimulated. Thus, for example, a blow to the eye, which has nothing whatsoever to do with optical phenomena, causes the recipient to "see stars." Obviously, the eye is not reporting accurately on the external world, for the reality is the blow, not the stars. How, then, is it possible to have confidence in what the senses report about the external world? Helmholtz examined this question exhaustively in both his work on optics and in his masterly *On the Sensation of Tone As a Physiological Basis for the Theory of Music* (1863). What he tried to do, without complete success, was to trace sensations through the sensory nerves and anatomical structures (such as the inner ear) to the brain in the hope of laying bare the complete mechanism of sensation. This task, it might be noted, has not been completed, and physiologists are still engaged in solving the mystery of how the mind knows anything about the outside world.

Helmholtz' detailed investigation of vision permitted him to refute Kant's theory of space by showing exactly how the sense of vision created the idea of space. Space, according to Helmholtz, was a learned, not an inherent, concept. Moreover, Helmholtz also attacked Kant's insistence that space was necessarily three-dimensional because that was how the mind had to conceive it. Using his considerable mathematical talents, he investigated the properties of non-Euclidean space and showed that these could be conceived and worked with as easily as the geometry of three dimensions.

Helmholtz' mathematical talents were not restricted to such theoretical planes as non-Euclidean geometry. He attacked and solved equations that had long frustrated physicists and mathematicians. In 1858 he published the paper "On the Integrals of Hydrodynamic Equations to Which Vortex Motions Conform." This was not only a mathematical tour de force, but, for a brief time, it also seemed to provide a key to the fundamental structure of matter. One of the consequences that flowed from Helmholtz' mathematical analysis was that vortices of an ideal fluid were amazingly stable; they could collide elastically with one another, intertwine to form complex knotlike structures, and undergo tensions and compressions, all without losing their identities. In 1866 William Thomson (later Lord Kelvin) proposed that these vortices, if composed of the ether that was presumed to be the basis for optical, electrical, and magnetic phenomena, could act exactly like primeval atoms of solid matter. Thus the ether would become the only substance in the cosmos, and all physical phenomena could be accounted for in terms of its static and dynamic properties.

Later life. Helmholtz' work in electricity and magnetism revealed his conviction that classical mechanics was probably the best mode of scientific reasoning. He was one of the first German scientists to appreciate the work in electrodynamics of the British scientists Michael Faraday and James Clerk Maxwell. Faraday had appeared to strike at the foundation of Newtonian physics by his unorthodox rejection of action at a distance, that is, action between two bodies in space without alteration of the medium between them. Maxwell, however, by interpreting the mathematics of Faraday's laws, showed there was no contradiction between Newtonian physics and classical mechanics. Helmholtz further developed the mathematics of electrodynamics. He spent his last years unsuccessfully trying to reduce all of electrodynamics to a minimum set of mathematical principles, an attempt in which he had to rely increasingly on the mechanical properties of the ether thought to pervade all space.

The mathematics of electrodynamics

Helmholtz was not in complete accord with Maxwell on the nature of electricity. Unlike Maxwell, Helmholtz was interested in and had studied electrochemistry, particularly the nature of the galvanic cell. Maxwell would have made the electric current solely the result of the polarization of the ether, or of whatever medium the current flowed through. Helmholtz, on the other hand, was fully conversant with Faraday's laws of electrolysis, which related the amount of current that passed through an electrochemical cell to the equivalent weights of the elements deposited at the poles. In 1881, in a lecture delivered in Faraday's honour in London, Helmholtz argued that if scientists accepted the existence of chemical atoms, as most chemists of the time did, then Faraday's laws necessarily implied

the particulate nature of electricity. This hypothetical particle was soon christened the electron and, ironically, the physics of its existence helped to falsify Helmholtz' theories of electrodynamics.

Though he was unsuccessful in his goal to formulate electrodynamics, Helmholtz was almost able to deduce all electromagnetic effects from the ether's supposed properties. The discovery of radio waves by his pupil Heinrich Hertz in 1888 was viewed as the experimental confirmation of the theories of Faraday, Maxwell, and Helmholtz. The special and general theories of relativity, proposed by Albert Einstein, destroyed Helmholtz' theories by eliminating the ether.

Helmholtz' early work on sound and music had led him to the study of wave motion. His work on the conservation of energy familiarized him with the problems of energy transfer. These two areas coalesced in his later years in his studies of meteorology, but the phenomena were so complex that he could do little more than point the way to future areas of research.

Helmholtz' work was the end product of the development of classical mechanics. He pushed it as far as it could go. When Helmholtz died in Berlin on September 8, 1894, the world of physics was poised on the brink of revolution. The discovery of X-rays, radioactivity, and relativity led to a new kind of physics in which Helmholtz' achievements, although impressive, had little to offer the new generation.

BIBLIOGRAPHY. There are two biographies of Helmholtz available to the reader of English. LEO KOENIGSBERGER, *Hermann von Helmholtz,* 3 vol. (1902–03; abridged Eng. trans., 1906, reprinted 1965), is often technical and sometimes difficult to understand. JOHN G. M'KENDRICK, *Hermann Ludwig Ferdinand von Helmholtz* (1899), deals only with Helmholtz' medical career. RICHARD M. and ROSLYN P. WARREN have published a collection of Helmholtz' writings on perception, entitled *Helmholtz on Perception: Its Physiology and Development* (1968), with critical comments. Helmholtz' *Popular Lectures on Scientific Subjects,* 2 vol. (1873; 2nd series, 1881), are excellent introductions to his thought.

(L.P.W.)

Heraldry

Heraldry is the science and art that deals with the use, display, and regulation of hereditary symbols employed to distinguish individuals, institutions, and corporations. These symbols, which probably originated as identification devices on shields, are called armorial bearings. Strictly defined, heraldry denotes that which pertains to the office and duty of a herald; that part of his work dealing with armorial bearings is properly termed armory. But in general usage, the term heraldry has come to mean the same as armory.

Like all other human creations, heraldic art has reflected the changes of fashion. As heraldry advanced from its utilitarian usages, its artistic quality declined. In the 18th century, for example, heraldry described new arms in an absurdly obtuse manner and rendered them in an overly intricate style. It was not until the 20th century that heraldic art recovered and in many ways improved upon the originals. There were still, however, far too many drawings of poor quality emanating from official sources.

This article is divided into the following sections:

The scope of heraldry 503
- General considerations 503
- The chief components of armorial bearings 505
- The elements and grammar of heraldic design 507
  - The field or ground of the shield
  - The "charges" on the field
  - The nature and origins of heraldic terminology
- The reading of heraldry 509
- Manipulation of heraldic design 510
  - Cadency
  - Arms of women
  - Quarterings and marshalling
  - Arms of bastardy
  - Nonfamilial heraldry

Historical development of heraldry 513
- Early roots of heraldry 513
- Growth of heraldry after the 13th century 515
  - Writers on heraldry
  - Continental versus British heraldry
- 20th-century heraldry 517
- Uses of heraldry for study and verification 518

Bibliography 518

The scope of heraldry

GENERAL CONSIDERATIONS

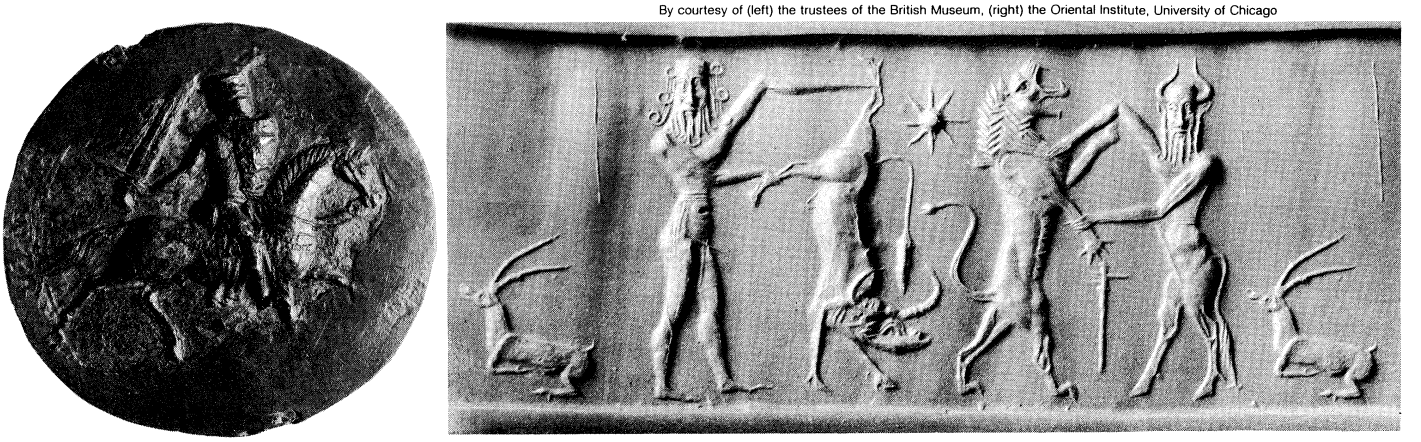
From the second quarter of the 12th century in western Europe, heraldic designs are found in general application. Elsewhere, a similar system is to be found only in Japan, in the *mon,* also dating from the 12th century. Other times and places are often said to have produced heraldic systems; for example, ancient Israel in the symbols of the 12 tribes, or the designs used by the Rajput princes in India. These and similar instances, however, are more properly considered incipient heraldry, since they did not develop into the complex heraldic practice known in western Europe and Japan.

From 1150 to 1500, the use of heraldry in the West was utilitarian: on armour in warfare, and on seals in peace. In the latter part of that period, it was used in peaceful ways and had much artistic value. Also, because from the beginning the use of arms had been associated with the higher feudal castes, heraldry acquired in later medieval times an identification with the concept of gentility that has persisted. To bear arms was the mark of a gentleman; therefore, to possess the desirable quality of gentility, a man needed to have armorial bearings. The great majority of those who seek to use coats of arms in the late 20th century are actuated by this motive. In the use of corporate arms, the motive of prestige rather than social distinction operates. As long as the possession of arms confers any social distinction, arms will be sought and used. At no previous time has there been so widespread an employment of heraldic devices.

Seals as ancient devices of heraldry

The use of symbols has been universal among civilized communities, but these symbols have not assumed the character always associated with heraldry. Seals, too, which have a prominent place in heraldic practice, are of an antiquity approaching that of the most ancient civilizations. They were in use in the states that from Sumer onward flourished in Mesopotamia. Their use, for example, in the Babylonian Empire was the same as in medieval western Europe: to authenticate the documents (possibly of baked brick, later papyrus, later still parchment or vellum) on which they appeared or to which they were appended. All persons, literate and illiterate alike, were able to recognize the representation or symbol of a ruler or other potentate. In 12th-century Europe, heraldry first appeared on seals in the representations of persons. There is a clear line of

By courtesy of (left) the trustees of the British Museum, (right) the Oriental Institute, University of Chicago



Seals as heraldic prototypes.
(Left) Impression of the great seal of Richard I of England, showing the mounted king bearing arms. In the British Museum. (Right) Cylinder seal impression from the Akkadian Period with a combat scene between a bearded hero and a bull-man, and beasts. In the Oriental Institute, University of Chicago.

descent from the seals of Assyria and Babylonia to the modern company seal, which is often heraldic.

Although originating in the small half continent of western Europe, heraldry has become universal, often, but not only, by way of western European colonization. Heraldry has spread to a considerable degree in both the Americas, Australia, New Zealand, and South Africa. In the former British India, the hereditary princes adopted the use of heraldry. In the numerous independent states formed in Africa from the former British colonies, official armorial bearings are generally used, and the same is true of the new states that were formerly French colonies. In Russia in the 18th century, the use of armorial bearings was adopted from the West, and state emblems are not unknown in Communist eastern Europe. In the 13th century, the Celtic princes of Wales and Ireland and the chiefs of the Scottish Highland clans took up the use of heraldic symbols from the example of the feudal lords and knights of other parts of Europe.

Other kinds of emblematic identification have some similarities with heraldry. An example is the totem system, found among the indigenous peoples of America and Australia, in which an animal, plant, or other object serves as an emblem of family or clan and is often regarded as a reminder of its ancestry. Totemism varies greatly in different countries, as do the theories that have been advanced to explain it. The totem poles used by the Indians of the northwest coast of North America contain a heraldic element in their employment of a hereditary symbol for a family or tribe. They therefore come under the heading of approaches to heraldic designs and may be termed semi-heraldic in character.

The Japanese *mon* as a type of armorial bearing

The Japanese *mon* is very definitely a heraldic symbol, having many parallels in its use with the armorial bearings of Europe. It was used on helmets, shields, and breastplates but never, as in Europe, large enough to identify the wearer of the armour at any considerable distance. When identification was desired, the *mon* was displayed on flags. The *mon* has usually been equated in English with "crest" and in some European languages has been translated erroneously as "coat of arms." It most closely resembles the heraldic badge (distinctive mark used by retainers), however, which in Europe often antedated armorial bearings. Further resemblances to European heraldry in the use of the *mon* include: the decorative use of the symbol on clothes, furniture, and houses; the use on the clothes of retainers of great lords; the legal requirement of registration of the *mon* (dating from the 17th century); and the reservation of the chrysanthemum *mon* to the emperor, with junior members of the imperial family using a different variety of the flower. This last distinction corresponds exactly to the rules of heraldic precedence that apply to the European royal families. That areas so far removed from each other as western Europe and Japan should have developed a system of hereditary symbolism independently of one another is not surprising, for in both areas feudalism was the prevailing medieval political and social system. As in Europe, Japanese heraldry survived the obsolescence of armour and has remained in widespread use in the 20th century.

Despite some uninformed opinion to the contrary, tartan has no connection with heraldry. It is simply a form of weaving cloth that is by no means restricted to the Scottish Highlands. Armorial bearings were adopted by Highland chiefs in imitation of the Lowland chivalry from the 13th and 14th centuries. The badge of the chief was adopted and used extensively by the members of his clan.

Drawing by Wm. A. Norman



The 16-petalled chrysanthemum *mon* of the Japanese Imperial family.

Flags can be heraldic. That of the United Kingdom is certainly so; it is formed by the amalgamation of the flags of England, Scotland, and Ireland, these showing respectively the crosses of St. George, St. Andrew, and St. Patrick, all of which are displayed heraldically. The United States flag has a quasi-heraldic character and appears to owe its principal ingredients to the armorial bearings of the first president, George Washington. The flag representing the republic of France, by contrast, is not heraldic, being merely an arrangement of the national colours.

In addition to national flags, there are banners, rectangular pieces of cloth showing the armorial bearings of the owner, and standards, strips of cloth that taper gradually to the end and usually bear heraldic devices but not the owner's full coat of arms.

Institutional arms from medieval times

An early development was the extension of heraldic design from its use by persons or families to its employment by institutions and associations of various kinds, a consequence of the concept that an assembly or body of people can be personified as an individual, much as a limited company or corporation is viewed as a legal "person." Medieval times provided numerous examples of arms borne by municipalities, churches, and colleges. The arms assumed by an individual or granted to him are regarded as being peculiarly his possession; therefore caution must be used in speaking of family arms. This question can be best dealt with in connection with the royal arms of the sovereign of the United Kingdom.

These arms are borne in their entirety only by the reigning king or queen. No other member of the royal family is permitted to bear the arms without introducing a "difference" mark that will show without doubt that the bearer is not the reigning sovereign. By analogy, the same condition holds for all so-called family arms, which belong to the head of the family; all other members should strictly bear them differenced—that is, with some mark of cadency (a sign indicating the position of the bearer with respect to the head of the family). In Scottish heraldry this rule is very rigidly enforced, but in England and elsewhere it has been allowed to fall into decay, except in the case of the royal family.

Probably the next development in the scope of heraldry was its use by ecclesiastics. The bishops and the abbots of the monasteries used arms on their seals from the 12th century onward. In this variety of heraldic usage, the arms were not those of individuals but of the body they temporarily represented—as also with arms borne by political units such as nations and cities or by educational establishments, many of which date from the Middle Ages. A great extension of medieval heraldry was connected with what came to be called the livery companies. These were guilds or associations of men in trades whose object was to uphold standards of craftsmanship. Most of them obtained charters from the crown and were granted arms. Among numerous examples in Britain are the Grocers, the Mercers, and the Glaziers companies. Membership in these still-existing companies no longer entails practice of their particular trades, but they possess property and have great charitable interests as well as considerable social esteem. Their armorial bearings are of great antiquity and are much displayed on their halls, letterheads, glass, silver, and so forth. Obviously, armorial bearings were assumed in the Middle Ages by such military bodies as the Knights Templars, the Knights of St. John of Jerusalem, the Teutonic Knights, and the great Spanish orders. Military heraldry has continued to the present: each of the three British armed forces, for example, has badges, or in some cases coats of arms, which are in the care of officers of the English College of Arms. The newest of the British armed forces, the Royal Air Force, alone makes use of more than 1,000 coats of arms or badges.

In the 20th century the development of corporate heraldry has gone far beyond anything known before. Throughout the world, banks, insurance companies, and many other great commercial concerns use arms, as do an ever-increasing number of professional, educational, and trade associations.



The arms of U.S. Pres. John F. Kennedy. The gold helmets are a variant on the three silver helmets of an ancient Kennedy coat. The border was added as a further difference. The olive branches and sheaf of arrows are derived from those of the Great Seal of the United States. The lambrequin is blazoned as argent and gules; this is exceptional to the rule that the principal tinctures of the arms (in this case or and sable) be repeated. No motto was included in the grant. The diagonal orientation of the shield is called couché and is optional in all depictions of arms.
Drawing by Wm. A. Norman

Perhaps the event most illuminative of the modern scope of heraldry was the grant, in 1961, by the government of the Republic of Ireland of armorial bearings to the president of the United States of America, John F. Kennedy. Because arms are hereditary and their owners are regarded heraldically as of noble status, the grant amounted to a bestowal of nobility by a state on the head of another state, an occurrence unique in heraldic history.

THE CHIEF COMPONENTS OF ARMORIAL BEARINGS

Heraldry originated when most men were illiterate but could easily recognize a bold, striking, and simple design. The use of heraldry in medieval warfare enabled combatants to distinguish one mail-clad knight from another and thus to know friend from foe. Thus, simplicity was the principal characteristic of medieval heraldry. In the tournament there was a more elaborate form of heraldic design. When heraldry was no longer used in war and heraldic devices had become a part of civilian life, an intricate type of design evolved with an esoteric significance utterly at variance with its original purpose. In modern times, heraldry has often been looked on as mysterious and a matter for experts only. Indeed, over the centuries its language has become intricate and pedantic. Such intricacy appears ridiculous when it is remembered that in the earlier periods swift recognition of a coat of arms or badge could mean the difference between safety and death, and in many medieval instances battles were lost through a mistake over the sameness of two devices of opposing sides.

The trend toward complexity and esoteric significance

The shield. The shield is the essential part of the armorial bearings; without it there can be no heraldic device, except for a woman, a distinction that calls for special treatment and is dealt with later. The word shield can be used to describe the coat of arms but in modern times is seldom employed in this way, except in a poetic context. Armorial bearings are generally referred to more briefly as "arms," or as a "coat of arms," a term derived from the surcoat of silk or linen worn over the armour to keep off the rays of the Sun and to prevent rust from forming on the armour. The surcoat was a waistcoat-like garment on which were shown the same heraldic insignia as on the shield.

Every other object in heraldic achievement is dependent upon the shield or coat of arms. There can be, and quite often is, a coat of arms consisting solely of a shield without any other object, such as a crest surmounted. The arms of the Churchills of Muston, a branch of the same family to which Sir Winston Churchill belonged, have no crest. The reason is that such families possessed arms before crests became fashionable.

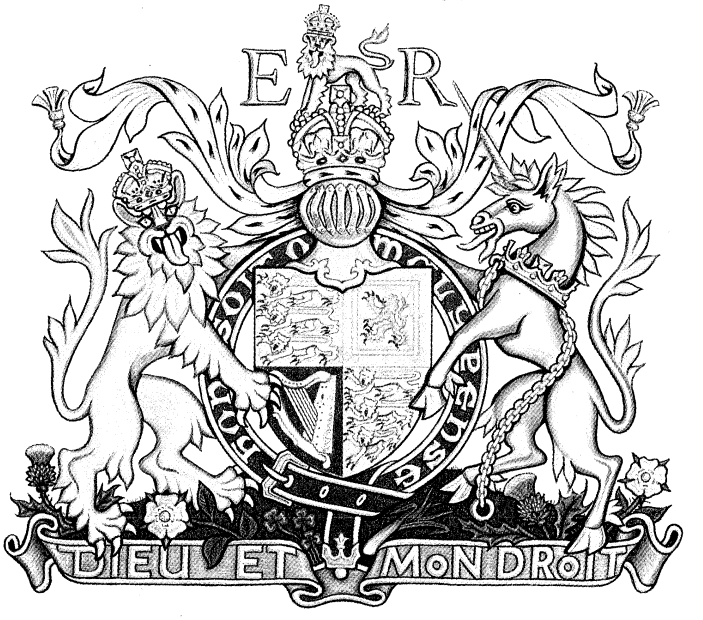
The crest. A crest is the object placed on top of the helmet and bound onto it by what is called the wreath of the colours, which shows the two main colours of the shield (see illustration of the royal arms of the sovereign of the United Kingdom). Crests were at first made of leather, later of light wood, and, as time went on, of more valuable materials. It is supposed that they were at first borne in tournaments; they became general in families in England from the 16th century when the venal heralds of that time persuaded crestless families to acquire the addition for a payment. Nowadays, a crest is automatically included in any grant of arms made in England, Scotland, or Ireland.

When horsedrawn carriages were in use, it was the rule to show the whole heraldic insignia on the coach or carriage door. With the advent of motorcars and their smaller door space, the arms were usually left off and only the crest and motto shown. This development may be the reason for the mistake frequently encountered in which the whole coat of arms is referred to as a "crest." It should be emphasized that while a coat of arms can exist without a crest, the existence of a crest without a coat of arms is an impossibility.

The crest and the age of the automobile

The helmet. On top of the shield is placed the helmet, upon which the crest is tied by the wreath. Originally, everything in heraldry was strictly utilitarian. As armorial bearings were used with the suit of armour, there had to be a helmet. In later centuries, rules for the depiction of

Drawing by Wm. A. Norman



The chief components of armorial bearings as indicated on the royal arms of the United Kingdom of Great Britain and Northern Ireland. The royal cipher (ER) is not a part of the arms proper but identifies them as representing Queen Elizabeth II. The roman numeral II is unnecessary here, as the arms of Elizabeth I were different, apart from those of England. The shield shows England (gules three leopards or) quartered with Scotland (or a lion rampant within a double tressure flory counterflory gules) and Ireland (azure a harp or stringed argent). This is the quartering in use since the accession of Queen Victoria in 1837. The shield is encircled by the garter of the Order of the Garter bearing the motto of the order, *Honi soit qui mal y pense.* The dexter supporter, a royally crowned gold lion guardant, and the sinister supporter, a silver unicorn with gold horn, hooves, mane, and tufts and a gold coronet collar and chain, represent England and Scotland, respectively. Atop the full-faced helm of a sovereign with its ermine and gold mantling, or lambrequin, is the royal crown surmounted by the royal crest, a lion statant guardant crowned with the royal crown. The motto *Dieu et mon droit* ("God and my right"), first used by Richard I, appears on the scroll below. The ground beneath the full achievement, called the compartment, is strewn with the floral and plant badges of England (rose), Scotland (thistle), Ireland (shamrock), and Wales (leek).

the helmet were elaborated to show the rank of the bearer, with some displayed in profile and some in full face, and with different metals and accoutrements to indicate status. The shape of the helmet has varied greatly in heraldic representation. While the basic features of heraldry remain unchanged, the modes in which the insignia are shown are subject to change and to fashion. The barrel-shaped helmet was used in the 13th century. The tournament helmet, often shown in drawings, was of a different type altogether, its shape resembling that of a soup tureen.

Mantling. From the helmet hung the mantling or lambrequin. This was of linen or other material and performed the useful function of shielding the wearer from the Sun's rays and also served to catch or deflect sword cuts. The mantling or mantle is painted with the principal colour of the shield, while its lining is of the principal metal. More elaborately styled mantles are used for kings and princes.

Crowns and coronets. These are emblems of the rank of the bearer. With the abolition of most of the great European monarchies, the study of crowns has become mainly of historical and antiquarian interest. The most famous royal crown remaining in the 20th century is that of the United Kingdom; it appears in the sovereign's arms upon the royal helmet and the crest of a golden lion crowned. Coronets (small crowns implying dignity inferior to that of the sovereign) are emblems of rank that are shown, when depicted, between shield and helmet. In Britain there are different coronets specified for the ranks of baron, viscount, earl, marquess, and duke. On the European continent, a much wider use of coronets has prevailed. Among the relics of this usage is the crest coronet, a coronet that supports the crest either instead of the wreath or in addition to it and resting upon it. Another is the chapeau or cap of maintenance, a cap with fur lining that was once worn on the helmet before the development of mantling and that is often used instead of the wreath to support the crest.

The myth of battle cries as sources of mottoes

Mottoes. Many myths have grown up around mottoes. They are often said to have originated as battle cries, but very few actually did. Among those that did are the *Crom a boo,* of the FitzGeralds, the dukes of Leinster, meaning "Crom [one of the family's old castles] forever"; and the "furth fortune and fill the fetters" of some Scottish families, which defies explanation and must refer to some forgotten incident. In succeeding centuries large numbers of mottoes were adopted. They are not part of a coat of arms and can be varied at the user's pleasure, though they are included in a modern grant of arms. More than one motto may be used by the same family. In Scottish arms the motto is usually shown above the crest, in all other countries beneath the arms and always contained in a scroll.

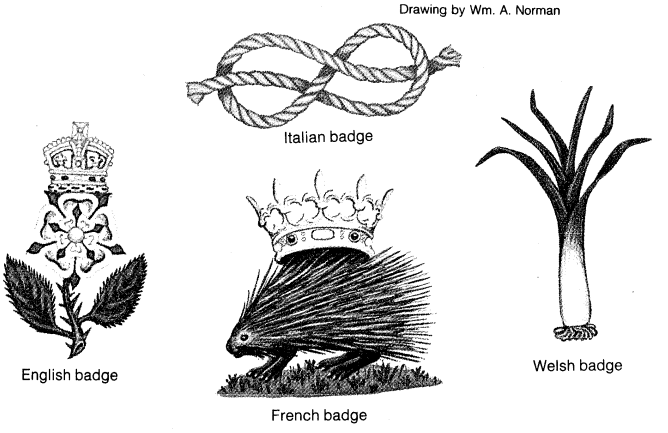
The supporters. These are the figures on either side of the shield of arms and are borne (in English heraldry) by peers, and by other bearers of orders of the highest class, such as Knights of the Garter, the Thistle, St. Patrick, and by Knights Grand Cross. In former times, supporters were used more widely, and a few English families still claim the right. In Scotland their use is much more frequent.

The compartment. The ground or foundation on which the supporters stand is called the compartment. In Scottish arms it is usually a rock or piece of ground strewn with some heraldic object. In England the compartment ought to be shown in the same way, and in the 20th century often is, with the scroll of the motto beneath it; but in the debased heraldic art of the 18th and 19th centuries, the supporters were generally shown as standing on a piece of ironwork or on the scroll.

The achievement. In heraldic writing, the term achievement often carries the same meaning as "arms," but probably its better usage is to describe the whole representation showing shield, helmet, crest, mantling, and supporters. The achievement belongs to only a minority of those who possess arms, since only a few have supporters. In addition, an achievement may include representation of various knightly orders or companionships of knightly orders to which the owner of the arms is entitled. For example, Viscount Montgomery of Alamein, British field marshal and World War II military leader, could show the symbol of the Order of the Garter around his shield; persons with lesser distinctions such as the Distinguished Service Order, Military Cross, and Order of the British Empire may have the decorations shown pendent from their shields. As distinctions of this kind are not hereditary, on the death of the bearer the successor to the arms must not use representations that show these honours.

Badges as symbols of persons or nations

The badge. This is older than the heraldic system. Such symbols expressing a person, a body, or an impersonal idea are found from ancient times. The eagle of Rome was one of the state's symbols and was the special device of the legions. Many such symbols bring to mind the country they represent; *e.g.,* winged bulls with human faces at once recall Assyria. On Trajan's Column in Rome, devices sometimes bear resemblances to later heraldic designs. On Etruscan vases are seen what not unfairly could be called demi-boars or bulls' heads caboshed. Nearer to heraldic times, the planta genista, or broom plant, which gave its name to the Plantagenet dynasty of England between 1154 and 1485, was a badge of the counts of Anjou before that family had armorial bearings. With the growth of heraldry, badges naturally assumed a heraldic character. They could be varied at the will of the holder, who often had more than one. Badges persist to the present and sometimes accompany a grant of arms.



Badges.
English badge: the red rose of Lancaster charged with the white rose of York, surmounted by the royal crown. Italian badge: the knot of the royal house of Savoy. French badge: the porcupine of Orléans, first used by Louis XII; the crown is not always included. Welsh badge: the leek; the daffodil is also a long-established badge of Wales.

Banners and standards. The users of arms in the Middle Ages often displayed them on the fork-tailed pennons of their lances. When the forked ends were cut off, the resulting flag was square, becoming a banner. Particularly valorous conduct was often indicated in this way, and the knight thus distinguished was known as a knight banneret. This last word was sometimes corrupted into baronet, a term that the English antiquary Sir Robert Cotton (1571–1631) used when he suggested to James I of England that he should revive a supposed order of baronets. The resulting Order (1611) was hereditary, however, and had no true connection with medieval knighthood. The banner bears the owner's arms, and in the 20th century anyone who possesses arms is entitled to a heraldic flag in this form. These can sometimes be seen in Great Britain flying over a house. No banner is referred to in the grant of arms made to President Kennedy, but in due course an armorial banner was made for the occasion when his brother, Sen. Robert Kennedy, carried it to the top of a mountain in the Yukon Territory named Mt. Kennedy by the Canadian government in memory of the President. The banner showed the Kennedy arms without the crest. The maker of the banner added a long white streamer on which he put the badge, the latter being part of the Kennedy crest.

On the standard, the main colours of the arms are shown with the owner's badge.

In England flags are often seen flying above churches.

When these show the flag of St. George (England's patron saint), white with a red cross, they carry in the top right-hand corner the arms of the diocese in which the church is situated. Heraldic flags also flew in countries on the Continent in a similar manner. With the disestablishment of heraldic offices in most European countries, contemporary flags are for the most part nonheraldic.

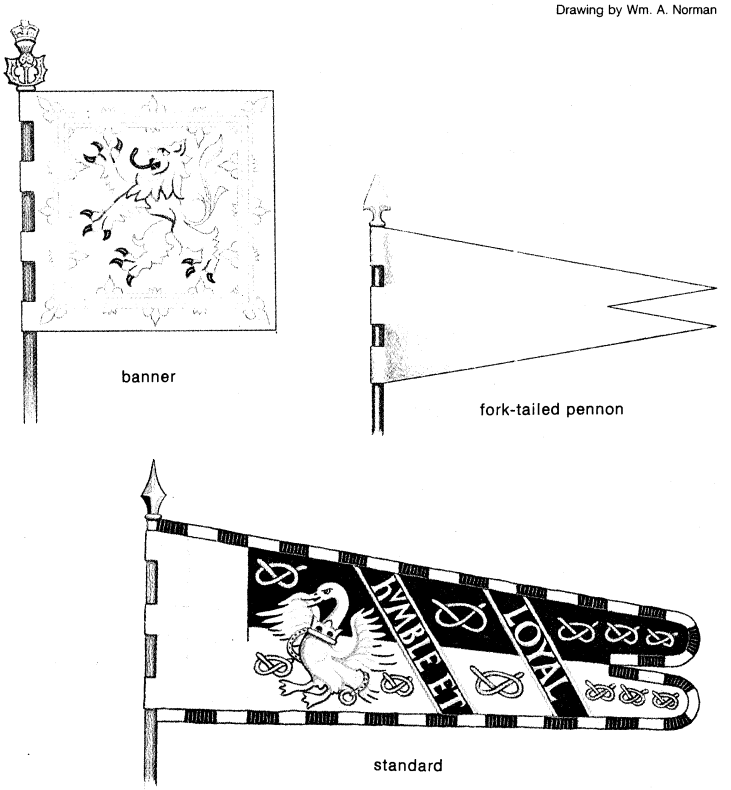
THE ELEMENTS AND GRAMMAR OF HERALDIC DESIGN

Provided that a few elementary principles are grasped, enough knowledge of heraldry can be acquired in a relatively short time to enable the student to understand the meaning of coats of arms. The multitude of terms used in heraldry need not worry him: once the rudiments are learned with some 50 of the terms, the meaning of the large remainder can be ascertained as the occasion arises. For example, when Queen Elizabeth II was crowned, some beautifully carved figures were made of the different badges that had been used by her ancestors, figures now displayed at Hampton Court Palace. They include one very rare badge—a yale. The yale is a mythical heraldic creature. Anyone unfamiliar with it could easily ascertain its meaning from the various heraldic glossaries. It is therefore unnecessary to burden the memory with hundreds of terms (a heraldic glossary generally contains about 800 terms).

Rudiments of heraldry in about 50 terms

The language of heraldry has a curious look. "Azure three wheat sheaves or" has been known to call forth the question, "Or what?" When it is remembered that *or* is the French for gold, the difficulty diminishes. Much heraldic terminology is a quasi-French, archaic language. In the Middle Ages, the French language was used by the ruling class in much of western Europe, so that it was not unnatural that heraldic terms should be French. In England, by about 1400, English words usually were used in preference. Much modern heraldic terminology, however, is so obscure that it seems purposely designed to puzzle the uninitiated.

Drawing by Wm. A. Norman



Heraldic flags.
Banner: the blazon of the shield is applied to the whole surface of a square or a vertically or horizontally oriented rectangular flag. This is the royal banner of Scotland, which follows the blazon of the second quarter of the royal arms of the United Kingdom. Although it is the banner of the sovereign, it is widely but incorrectly used today as the national symbol. Fork-tailed pennon: shown here is that of the Sovereign Military Order of Malta; gules a cross argent. Standard: the cross of St. George at the hoist identifies this as English. The profusion of badges, the diagonally placed motto, and the border of alternating tinctures are typical. This is the standard of Sir Henry Stafford, *c.* 1475.

The terms dexter and sinister mean merely "right" and "left." A shield is understood to be as if held by a user whom the beholder is facing. Thus the side of the shield facing the beholder's left is the dexter or right hand, and that opposite his right, the sinister or left hand.

The field or ground of the shield. The field or ground of the shield is of one of three kinds: a colour, a metal, or a fur. There are five main colours (known as tinctures): azure (blue), gules (red), sable (black), vert (green), and purpure (purple). In English heraldry there are also murrey (sanguine, a tint between gules and purpure), and tenné (an orange-tawny colour). The metals are or (gold) and argent (silver, which is often shown in coloured illustrations as white). The furs are ermine (white field with black spots), ermines (black field with white spots), erminois (gold field with black spots), pean (black field with gold spots), and vair (composed originally of pieces of fur from a species of squirrel that was blue-gray on the back and white underneath, so that when several of its skins were sewn together, the result was a series of cup-shaped figures, alternately blue and white, which is the way vair is shown).

The furs, from ermine to vair

It is considered bad practice to put a colour upon a colour, a metal upon a metal, or a fur upon a fur. Many examples of such bad heraldry are found in old records, but in this as in other instances, the rules now prevalent grew up only very gradually.

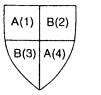
The "charges" on the field. The field is said to be "charged" with an object. Heraldic objects are of a large and growing variety; as more and more arms are devised, more and more objects appear as charges—telescopes, aircraft, rolls of newsprint, and so on. Charges are divisible into two classes: the ordinaries (or honourable ordinaries as they are often called, deriving their name from the fact that they are so frequently used) and the others.

The ordinaries. The ordinaries comprise some 20 figures. Among them are: the chief, being a third of the shield and the top part; the pale, a third part of the shield, drawn perpendicularly; the bend, a third part of the shield, drawn from the dexter chief to sinister base; the bend, sinister, drawn from the dexter base to sinister chief; the fess, a third part and taking up the centre of the shield; and the chevron, resembling an inverted stripe in the rank badge of a noncommissioned officer. It should be noted that the bar is a diminutive of the fess, of the same shape, and can be placed in any part of the shield. The term bar sinister is often used in fiction as a synonym for bastardy. It has no such significance, bastardy being denoted heraldically in several other different ways. Since the European nations were Christian when heraldry was invented, the cross appears in many forms in heraldry. The cross in a coat of arms does not imply, however, that the original bearers were crusaders.

The border, or bordure, is used as a mark of difference sometimes, and in English heraldry since the mid-18th century it is used as a sign of bastardy. The orle is an inner border, not touching the ends of the shield; the field is seen within and around the orle, giving it the appearance of a shield with the middle cut out (voided, in heraldry). The tressure, much used in Scottish heraldry, is a diminutive of the orle. The inescutcheon, a small figure shaped like a shield in the middle of the shield, is used to denote the arms of a heraldic heiress; the quarter occupies one-fourth of the shield; the canton, less than the quarter and one-third of the chief; and checky or chequy, the field or charge divided by transverse lines horizontally or perpendicularly into equal parts. Billets are oblong figures. If their number exceeds 10 and they are irregularly placed, the field is described as *billetté;* the paile or pall is the upper half of a saltire (St. Andrew's Cross) and half a pale. The pile is in the shape of an inverted pyramid. The flanch, or flanque, is a segment of a circle drawn from a corner of the chief to the base point. The lozenge is a parallelogram having equal sides forming two acute and two obtuse angles, and a mascle is a lozenge voided. The roundel is circular in form. Lozengy is the field divided

by diagonal lines transversely, and the fret is a mascle interlaced with a saltire.

Minor charges. A field is said to be "powdered" or "semé" when strewn with minor charges; when charged with drops of liquid, it is "gutté." Partition lines divide the shield. The most common ones are straight. "Impalement" or "dimidiation" means the division of the shield into two equal parts by a straight line from the top to bottom. This method is used to show either the arms of husband and wife, the arms of the husband being in the dexter half, or certain types of official arms, as the arms of a bishop's see impaled with his family arms, those of the see being in the dexter half. The shield is divided into four quarters when one coat of arms is quartered with another, as when the children of a heraldic heiress (the eldest daughter of a family of no sons) use their mother's arms with their father's. (Thus in the illustration A = father's arms, B = mother's arms.)



two coats of arms quartered

Other divisions of a shield are: party per pale or simply per pale, division of the field into two equal parts by a perpendicular line (this resembles the impalement just mentioned but does not serve the same purpose of combining arms); party per fess, division into two equal parts by a horizontal line; party per bend; party per chevron; party per saltire; gyronny of eight. When the partition lines are not straight, they can be of several varieties.

The nature and origins of heraldic terminology. Fanciful explanations have been advanced to account for heraldic charges: for example, argent to denote purity, the bend derived from the military cross belt—the cross a sign of a crusading ancestor—and so on. Since no one wrote about heraldry until it had existed for more than 200 years, these explanations of its symbolism can be discounted. With very few exceptions, the origin of the charges is unknown. The Stourton arms ("sable a bend or between six fountains") refer to the six springs in the park of their ancestral estate that are the source of the river Stour. A heraldic fountain does not resemble a real fountain but is put in the form of "roundel wavy argent and azure" (a silver and blue circlet of wavy lines), unless it is expressly stated that the fountain is "proper"; *i.e.*, a natural fountain. The word proper is always used to denote a charge shown in its natural colours or natural form.

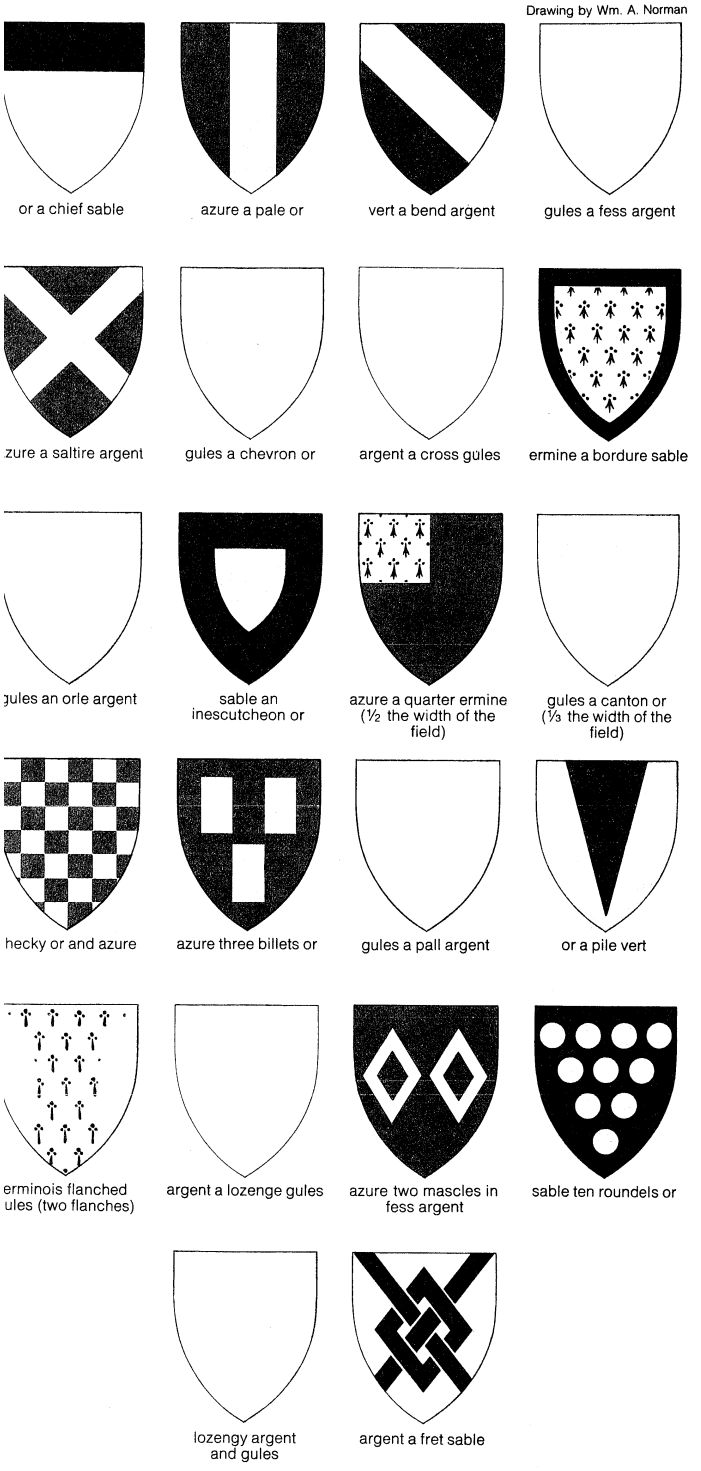
Augmentations of honour

The derivation of heraldic charges is more easily discerned in the augmentations of honour, as they are called, when something has been added to a coat of arms by the (British) crown in recognition of services rendered. The arms of the British naval hero Admiral Horatio Nelson show fresh heraldic charges added to his ancestral arms as his victories were gained. Within the past 300 years, augmentations have generally been recorded. An example is the augmentation granted by Queen Victoria to commemorate the discovery by the English explorer John Hanning Speke of the sources of the Nile. The honour, granted posthumously, consisted of the addition to the existing arms of a chief azure upon which appeared a representation of flowing water proper superinscribed with the word Nile in gold lettering. Usually, however, the origins of the various objects used in heraldry are not known. Numerous historical instances of augmentations of honour occurred in continental Europe, especially in connection with the Holy Roman emperors. Frederick II, for example, granted to Conrad Malaspina an augmentation of a chief of the empire, thereby adding an eagle displayed sable to the Malaspina arms of per fess gules and or overall a thorn branch vert with five flowers argent in pale.

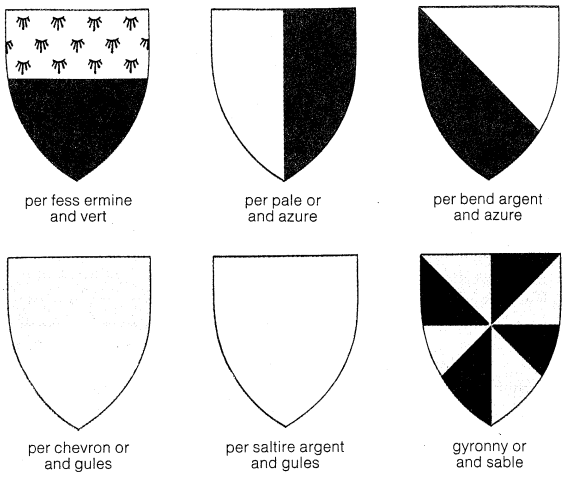
Heraldic descriptions are called blazons. The term is derived from the French *blason,* the etymology of which is uncertain. Originally, it denoted the shield of arms itself and still retains this meaning, but it is now generally used in a derivative sense as meaning the description of the arms. Blazon is thus a noun, and there is also the verb to blazon; *i.e.*, to describe a coat of arms.

There are four generalizations that are useful in the deciphering of blazons. First early coats of arms are simple because they were original and there were so few of them that elaborate differentiation was not required. As time brought many more coats of arms into being, simple coats became much rarer, and the passing of warlike usage made arms much more complicated. Second, punning or canting arms are very common as, for example, trumpets for Trumpington, or a spear for Shakespeare. It is notable, however, that many armorial allusions that were formerly obvious now require research for elucidation. Other allusions have been lost entirely. Third, in grants of arms to people bearing the same name but having no relationship with each other, difference marks have had to be put in.

Drawing by Wm. A. Norman



Ordinaries, basic bearings that may be of any tincture and that may be combined in great variety. A combination of a cross (England) and two saltires (Scotland and Ireland) has resulted in the familiar Union Jack of the United Kingdom. Ermine and certain other textures such as ermines (black with white ermine tails) are regarded as tinctures in their own right and may bear superimposed charges. Discrete charges such as lozenges, mascles, fleurs-de-lis, etc., may be used singly, in pairs, in threes or greater numbers, sometimes in great profusion, as that of ermine tails.



Partition of the shield.
The field is often divided along the lines occupied by ordinaries, just as quartering imitates a cross. "Per fess" means along the line over which a fess would be laid down. The ermine tails illustrated are one type of stylization among many in use. The superior dexter segment on the gyronny shield is called a gyron and is occasionally found singly.
Drawing by Wm. A Norman

Again, in consequence, blazons have become much more complicated. Finally, in the course of centuries and frequent intermarriages among arms bearers, many quarterly and grandquarterly coats have come about. Quarterly and grandquarterly coats are much more difficult to describe than the simple coats.

Apart from the ordinaries and those other charges that have been mentioned incidentally, there are some peculiarities of heraldic charges that need to be noted. Mythical birds and animals are much used, the product of ancient and medieval natural history—or the lack of it. Such are the dragon, griffin, wyvern, harpy, phoenix, and martlet. In addition, there are some creatures bearing the names of real animals but not resembling them in all respects. The heraldic tiger is more like a lion or a wolf in some features. When the real tiger became known to heraldry, it was described as a Bengal tiger. The heraldic description of animals is very important. Rampant means on the hindlegs, while rampant guardant is the same posture but full faced. Reguardant means looking back; passant, walking. Combattant signifies two animals fighting on hindlegs. Couchant is lying down; dormant, sleeping; and sejant, sitting. A beast of the hunt is called at gaze when looking full face, trippant when at trot with one foot raised, and statant when standing. Parts of an animal may be a charge; *e.g.*, a demilion or demiwolf, a lion's or bear's gamb or foreleg. Heads are described as erased when cut off by a jagged line, couped when cut by a straight line, and caboshed when the severed head looks straight forward. A bird shown with wings expanded is said to be displayed. Creatures placed back to back are addorsed. A fabulous bird, the phoenix, is known to heraldry; also known was the legendary pelican that fed her young on her own blood and was called "in her piety," then being considered an emblem of Jesus Christ, who fed or redeemed his flock with his own blood. The martlet is another fabulous bird widely known outside heraldry, owing to John Milton's reference to the herald's martlet, which has no legs. It is a frequent charge, resembling a swallow, and is used in cadency to denote the fourth son. Other terms have special heraldic significance. Armed is used of the horns, teeth, or claws of a beast, or the beak or talons of a bird, and of the human being when in armour. The term slipped applies to flowers and fruit when the stalk is seen. Counterchanged refers to the field when it has two tinctures, a metal and a colour, and when one is the background for the other on one side of the shield but the relationship is reversed on the other side. An example is the Warner arms: "per bend argent and gules two bendlets between six roses all counterchanged," where the three roses on argent will be gules and the three on the gules will be argent. The field is separated by a partition line and the charge or charges of the arms are said to be countercharged when the charge or portion of a charge that lies on the metal is of the colour and vice versa.

Heraldic descriptions of animals and their postures

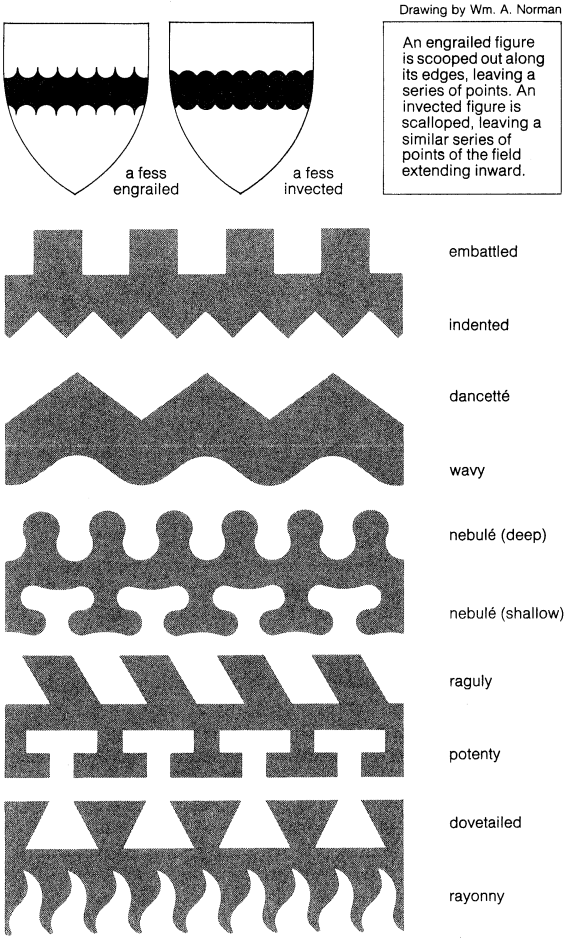
THE READING OF HERALDRY

The Petra-Sancta system of denoting colour by black-and-white patterns

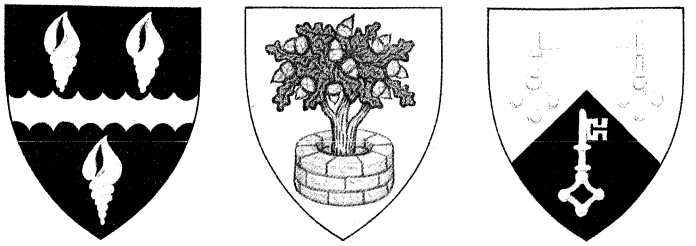
A method has been devised to indicate heraldic colours in black-and-white illustrations. Known as the system of Sylvester Petra-Sancta, an Italian herald, it makes use of the following equivalents: or is denoted by dots or points, azure by horizontal lines, vert by lines from dexter chief to sinister base, purpure by lines from sinister chief to dexter base, argent by a plain field, gules by perpendicular lines, and sable by cross lines horizontal and perpendicular. Furs are depicted with black or white spots on the appropriate ground; vair and countervair are shown by alternate lines and plain surfaces.

Describing, or blazoning, of arms must always begin with an identification of the ground of the shield, such as argent or gules or ermine. For a woman who is not married, the arms normally appear on a lozenge, not a shield, but the field or ground in this instance, too, must be the start of the blazon. Then come the charges. A typical blazon is thus: "sable [ground of shield] a chevron ermine between three lions rampant argent crowned or" (arms ascribed to a family of Hinstoke). The chevron is a fur; the lions are silver, appear on the sides of the chevron and its base, and have gold crowns. One important feature in heraldic writing is economy of words. Technically, it should be possible to avoid punctuation marks, thus: "azure a fess between three stags trippant or" (Hind). Here both fess and stags are in gold. When three beasts are depicted, they are shown in the most convenient way around the main charge; that is to say, two in the upper part of the shield

Drawing by Wm. A. Norman



Types of divisions between tinctures.
A line described as flory, or flory counterflory, employs a series of small fleurs-de-lis that have substance of their own beyond the two areas being divided. Rayonny (or *rayonné*) may alternate straight points with curved points.

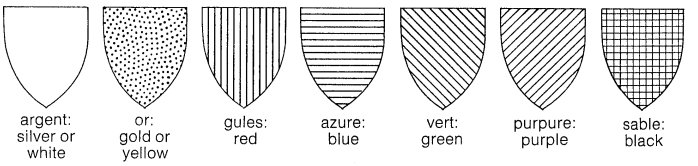


Shelley Wellwood Keyes

Canting, or punning, arms, derived from the literal meaning or from the sound of a name.
(Left) Shelley: sable a fess engrailed between three whelk shells or. (Centre) Wellwood: argent an oak tree growing out of a well all proper. (Right) Keyes: per chevron gules and sable, three keys or.
Drawing by Wm. A. Norman

and one below. A straightforward coat with only one charge on the field is that of the Italian Segni family of Agnani, which gave to the church the popes Innocent III, Gregory IX, and Alexander IV: "gules an eagle displayed chequy sable and or." Economy, however, can be carried too far as in the following: "azure a lion rampant double queued barry of ten argent and gules armed and langued of the last crowned or, within a bordure of the second and third" (Mountbatten). Here is an example of a usage that grew up in past centuries and was designed to avoid repetition of the name of a tincture but in reality served to create confusion. "Of the last" means that the lion's claws and tongue are in red or gules. "Of the second and third" means simply argent and gules. There is no real economy since more words actually are used and reference has been made to the earlier parts of the blazon. This type of blazon is used when a large amount of heraldic insignia has to be described, but it makes such long blazons unnecessarily complicated. Anyone who is writing a blazon should not use this jargon of last, first, second, and so on.

Drawing by Wm. A. Norman



Conventional representations of tinctures used by engravers and others when actual colours are not practicable.

Blazoning the crest

The helmet is the next item to be characterized, although in blazons it is usually taken for granted and left undescribed. When it is mentioned, it is said to be "befitting his degree." Although the helmet need not appear in written descriptions, it always should be depicted in illustrations. It is a bad feature of many drawings that the helmet is absent, showing the crest as if it were airborne above the shield and thus unsupported. The crest must always be mentioned when, as in the vast majority of cases, one exists. In formal blazons, the wreath (also called the torse) is given as well; thus, crest—"on a wreath of the colours, a wolf passant proper" (Trelawny). The wreath is not usually mentioned, however, because like the helmet it is always assumed to be there. The term colours refers to the two chief colours of the ground of the shield. As with the shield, the older the crest, the simpler it will be. Most people can envisage on a knight's helmet the figure of a wolf walking, but it is difficult to picture someone in armour wearing as his head crest "the stern of a Spanish man-of-war on waves of the sea all proper thereon inscribed 'San Josef' with the motto above, 'Faith and Works' " (Nelson). This latter example belongs to the period of decadent heraldry in the late 18th century and 19th century in England.

The mantling, or lambrequin, is mentioned in formal descriptions but not in general usage. The supporters and compartment pertain only to a few classes of arms bearers, and in descriptions the supporters are blazoned after the crest (or crests). The compartment is not usually described but sometimes has to be, as in the arms of the earl of Perth: supporters (two savages = two ancient Caledonians) stand on "a compartment strewn with caltraps" (from "caltrops," iron instruments designed to maim horses' feet and used by the Scots with great effect at the Battle of Bannockburn, 1314).

The motto comes at the end of the description, not being part of the arms. The badge is rarely found, except among very historic families (and by a strange inversion in some 20th-century grants), but when it occurs it, too, comes at the end of the blazon. It can be very simple, as with that of Lord Mowbray, Segrave, and Stourton—"a sledge or." It may be very elaborate, as with Constantine—"a hurt [*i.e.*, a roundel in azure] charged with a leopard's face and surmounted upon the edge with two fleurs-de-lis in pale or, and as many roses in fess, argent, barbed and seeded proper." In this example the roses are silver but the leaves are proper. Coronets of rank are not usually mentioned in English or Scottish heraldry, but caps of maintenance and crest coronets must be blazoned with the crest. Banners and standards are not as a rule mentioned in blazons, though they may be when they occur in a modern grant.

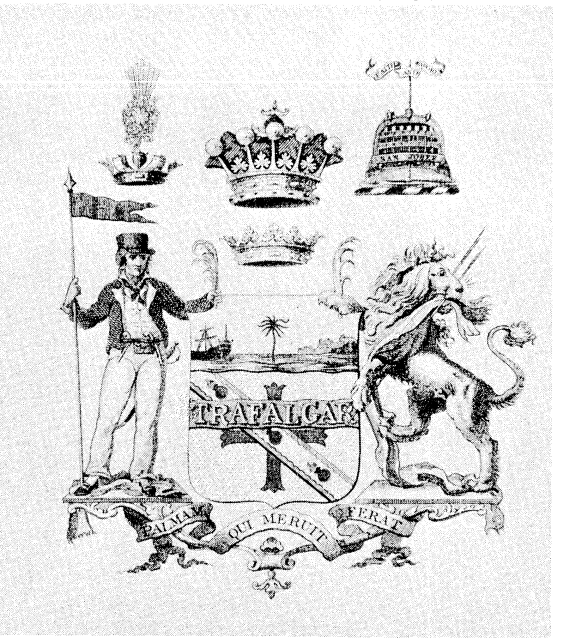
MANIPULATION OF HERALDIC DESIGN

It is clear that the vast majority of heraldic charges are without the foundation of legend often assigned to them, though in many instances the real origin of the charge is lost. A Jacobean dramatist could write of a family as old as the first virtue that merited an escutcheon—a nice poetic flourish that should remain in poetic realms. In modern grants the heralds try to give some allusion to the grantee's work and achievements and his place of origin, and canting arms still appear. Many arms are recondite, however, in their significance, and much has to be learned before the significance of the charges is known.

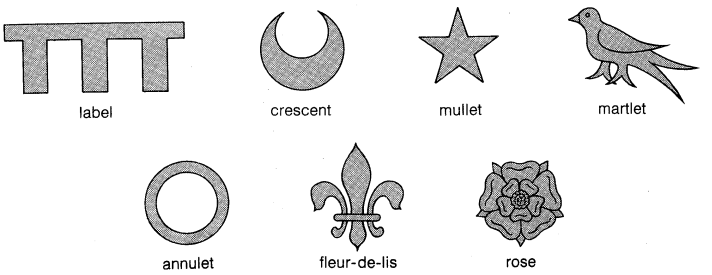
The use of devices to show position in family

Cadency. The rules evolved over the centuries to denote particular distinctions in heraldry are fairly straightforward. Cadency is the use of various devices designed to show a man's position in a family, with the aforementioned basic aim of reserving the entire arms to the head of the family and to differentiate the arms of the rest, who are the cadets, or younger members. Heraldic works in the 16th century refer to cadency marks as: a label for the eldest son during his father's lifetime; a crescent for the second son; a mullet (five-pointed star) for the third; a martlet (a mythical bird), the fourth; an annulet (a small ring), the fifth; a fleur-de-lis, the sixth; a rose, the seventh; and so forth. These marks were not always used in the Middle Ages. Differences might be shown instead by a

By courtesy of the National Maritime Museum London; photograph, Patrick Rossmore



Earl Nelson's coat of arms, drawn in sepia, 1806.



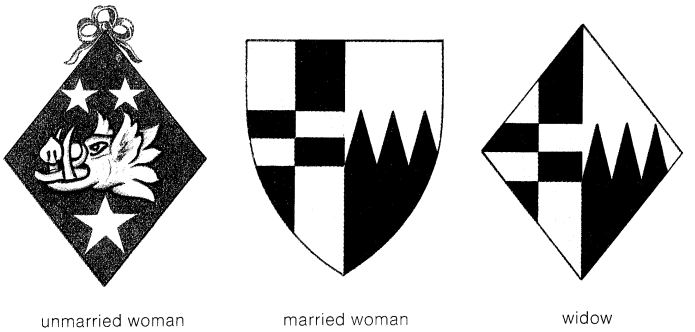
Marks of cadency, used to difference the arms of cadets of the same family. The label, the mark of the eldest male heir, is a notable feature of the arms of the prince of Wales, the heir to the throne. The second cadet displays a crescent, the third a mullet, and so on. These symbols may be of any tincture and may be used otherwise than as signs of cadency.

Drawing by Wm. A. Norman

change of tincture, by adding small charges to the field, and the like. Both on the Continent and in England, rules of cadency have long ceased to be used. It is customary for all members of a family to use the entire arms of the head. There are, however, two exceptions. Very occasionally, a crescent is used for difference by a noble family showing descent from a second son. The other exception occurs in the arms of the British royal family, in which the cadency system exists in rigour. The reason is that the royal arms are arms of sovereignty and cannot be shared. The sovereign alone can have the whole undifferenced arms. Nor does any member of the royal family—not even the prince of Wales—have any right to the use of arms until they have been granted to him by the sovereign. A label of difference with marks is placed on the arms; a three-pronged label for the children of the sovereign, a five-pronged label for grandchildren. The Duke of Windsor, after his abdication as Edward VIII in 1936, was granted arms with a label.

In Scotland the position on cadency is very different. Since heraldry is regulated in Scotland by acts of the Scots Parliament before the Union in 1707 with England, and confirmed by the British Parliament, the regulation of arms is very precise. The strict observance of cadency is probably because the Celtic clans formed the original social system in Scotland before the advent of feudalism. Thus only the chief of the name can have the entire arms. He matriculates, or enters his arms, in the registers of the Lord Lyon King of Arms (whose court has jurisdiction over armorial bearings in Scotland). This registration also applies to his eldest son (subject to suitable differencing of the arms) who inherits them in due course. The younger sons must petition for a matriculation of the paternal arms with a suitable difference indicating the position of each in the family. As families from the descendants of the original grantee continue to be established, so there is matriculation and rematriculation, in a carefully prescribed manner.

Drawing by Wm. A. Norman

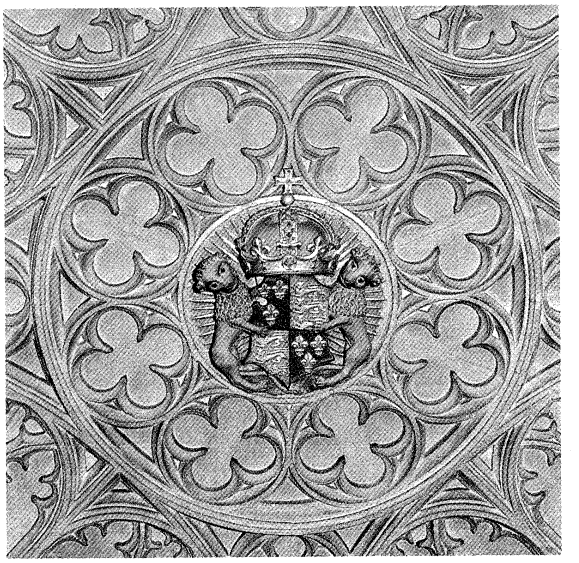


Arms of women.
A woman adopts the undifferenced arms of her father. An unmarried woman: the arms of her father are displayed on a lozenge; a true lover's knot signifies unmarried status. A married woman: the wife's arms, to the sinister, impale those of the husband, to the dexter; the husband displays the combined arms as head of the family, and the wife shares his escutcheon. A widow: the woman's arms revert to the lozenge, retaining the deceased husband's impaled arms.

Arms of women. Arms of women are shown during spinsterhood or widowhood on a lozenge, not on a shield, without a crest (except in Scotland, where a woman who is chief of a clan and head of the name, such as MacLeod of MacLeod, is allowed a crest). A woman divorced and not remarried also uses a lozenge. The arms of a married woman are shown in conjunction with her husband's by impalement, the division of the shield into two equal portions, the husband's arms on the dexter and the wife's on the sinister. Should she be a heraldic heiress, the arms of her family are placed upon an inescutcheon (originally inner escutcheon) or "escutcheon of pretence" (a small shield whose position in the fess point of the husband's shield gives it precedence over all of the other parts of the shield). The only exception to these rules for women is that of a queen regnant like Elizabeth II, who, being sovereign and thus considered in heraldic terms to be herself the source of honour for all her subjects, possesses the full arms of sovereignty of her royal house and kingdom.

In the Middle Ages, arms for women were often shown on shields, and those like Joan of Arc who bore arms in battle may have used crests.

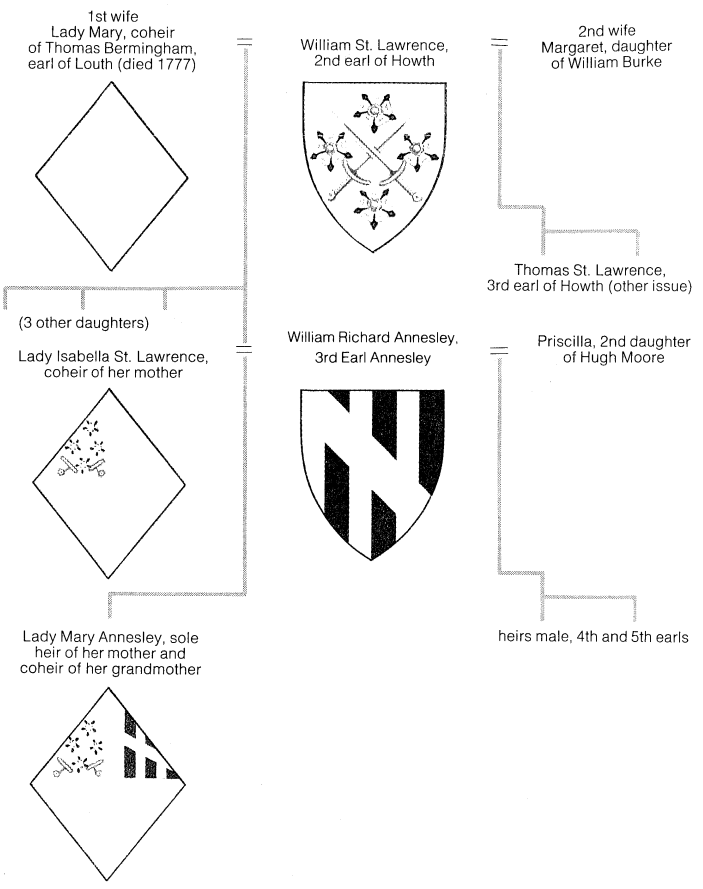
Picturepoin



Royal arms of England as used from Henry V's time, quarterly France and England, first used by Edward III but by him with the French quarter showing strewn fleurs-de-lis. Roof boss from St. George's Chapel, Windsor Castle, 1483–1528.

Quarterings and marshalling. In the quarterings and the marshalling arrangement of more than one coat of arms in the same shield, the position of heiresses must be considered first. The children of a heraldic heiress are entitled on her death to quarter her arms with their father's (the arrangement is to show the shield divided into four quarters so that quarters 1 and 4 are the father's arms, 2 and 3 the mother's). This positioning of the quarterings is also used in England when an additional surname and arms are taken, almost always in obedience to a will. This is the "name and arms" clause peculiar to English law. Its operation over the last 200 years is responsible for the double- or treble-barrelled surname common in England and also found in Scotland. Thus in Salusbury-Trelawny, the original Trelawny arms appear in quarters 1 and 4 and the assumed additional arms for Salusbury in 2 and 3. A famous historical case is that of King Edward III of England, who in 1340 claimed the throne of France in right of his mother, a French princess. He then quartered the lilies of France (the fleurs-de-lis) with the lions (leopards) of England. England should have been placed in 1 and 4, but Edward gave this position to France, probably because of the greater size and resources of France at that time. In this form, the royal arms continued until 1800, when the empty title of king of France was dropped and the lilies went out with it.

When quarterings are inherited from a woman, no crest is transmitted with them, because a woman cannot pass



Heirship through the female lines.
Drawing by Wm. A. Norman

on a crest. The matter alters, however, when additional arms are taken in obedience to a will; then a double crest is likely. There is no reason why a further assumption may not occur, so that triple or quadruple hyphenated names are found: for example, the English county family Sawbridge-Erle-Drax has quarterly arms, 1 and 4 Drax, 2 Erle, and 3 Sawbridge. This type of quartering is not difficult to follow, but a real problem in marshalling several coats in one shield arises when more than one heraldic heiress occurs in the same family. Some families of long descent have often married heraldic heiresses, acquiring many quarterings. Sometimes several hundred quarterings are attributed to the head of a great family. A splendid instance of quartering occurred in the achievement of the empress Maria Theresa. Before her accession to the imperial throne she was queen of Hungary and Bohemia and by marriage grand duchess of Tuscany. As a sovereign in her own right she bore a shield on which there were 29 quarterings. The dukes of Rohan-Chabot in France bore a quartered shield with Navarre, Scotland, Brittany, and Flanders; overall was an escutcheon of Rohan quartering Chabot.

Problem of marshalling several coats in one shield

Even without such large numbers of arms to deal with, the marshalling of quarterings is still a problem. Various methods have been tried, often with results as difficult to decipher. An interesting illustration of the marshalling of several coats of arms is that of the baronet Cameron-Ramsay-Fairfax-Lucy. The arms are said to be quarterly with the arms of Lucy in 1 and 4. Then in 2, the description runs, grand quarter counterquartered. This means that quarter 2 is itself a quarterly coat, 1 and 4 of which are for Fairfax, 2 for Ramsay, with the third quarter counterquartered; *i.e.,* itself quartered, showing two coats of arms. The third quarter is that for Cameron.

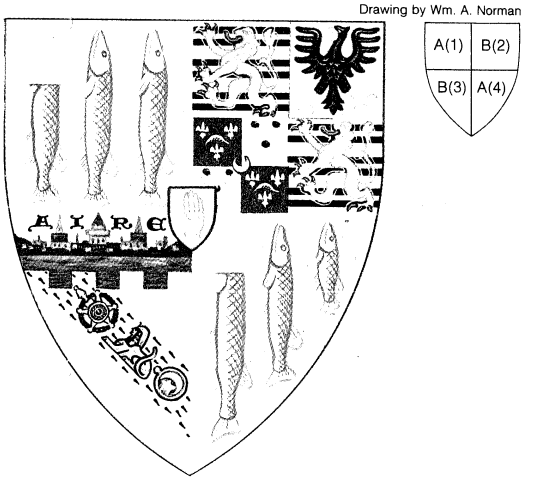
Arms of bastardy. The heraldic illustration of bastardy was achieved in several ways in the older days of chivalry. Little social or moral obloquy attended the status of bastardy, possibly because of the late marriages of the upper classes and their arranged unions. Thus the arms of a bastard were merely differentiated as perhaps those of a distant cadet line would be. From early times the bend or bendlet sinister was used. The erroneous use of bar sinister might have come about because the French for bend sinister was *une barre.* In the arms of British royal bastards like the dukes of St. Albans, of Buccleuch, of Grafton, and others, a baton sinister is used to denote bastardy. The royal arms of these illegitimate scions are said to be "debruised" (crossed or partly covered) by the baton sinister. English heralds in the last 150 years have often signified bastardy by the use of the bordure.

Nonfamilial heraldry. The arms of grantees other than families or individuals are often encountered. In modern times the granting of arms to private persons has ceased in some countries where grants of corporate arms are frequent. In Sweden, for example, grant of neither title nor arms occurs, but grant of arms to public bodies, such as local administrative units, is frequent. Historically, it was an easy passage from the arms of individuals to those of corporate bodies. This is particularly evident in the military sphere, where the great crusading orders led to the many important orders of chivalry in the principal European countries. The Elephant of Denmark, the Golden Fleece of Spain and of Austria, the Holy Spirit of France, the Garter of England, and the Thistle of Scotland were all preceded by the orders of military monks, and all have insignia that contain heraldic features and that occur in many arms illustrations. Most of the older bishops' sees have official arms; in the Anglican Church the missionary bishops as well as the diocesan bishops have arms. In the Roman Catholic Church the episcopal sees all have arms; and new arms are granted by the pope, who, as head of the Vatican state, is a temporal sovereign as well as spiritual head of the church. The arms of the popes often contain charges that are added to their individual arms after their election to the papacy or to earlier ecclesiastical office.

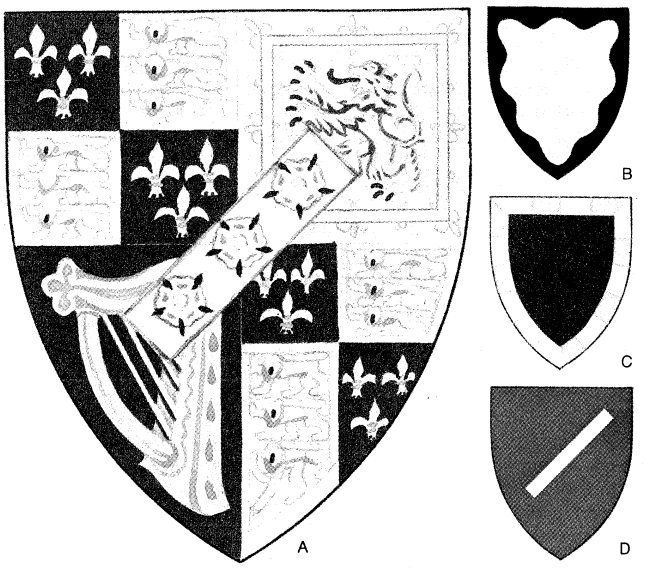
The military orders of chivalry

Dominion and colonial arms are necessarily interconnected with royal arms, since the British crown in each country is or was the source of honour and must have granted arms to its various territories. Because of the vast extent of the former British Empire, the richest collection

Drawing by Wm. A. Norman



Marshalling of several coats of arms.
(Left) The arms of the Cameron-Ramsay-Fairfax-Lucy family, blazoned: arms-quarterly, 1st and 4th gules semé of cross-crosslets, three lucies hauriant argent, a canton of the last (Lucy); 2nd, grand quarter counterquartered, 1st and 4th argent, three bars gemel sable surmounted of a lion rampant gules, armed and langued azure (Fairfax); 2nd parted per pale argent and or, an eagle displayed sable, armed beaked and membered gules (Ramsay); 3rd counter-quartered, 1st and 4th azure a branch of palm between three fleurs-de-lis or; 2nd and 3rd gules three annulets or stoned azure. In the centre of these quarters a crescent or (Montgomerie); 3rd gules, three bars or, on a bend ermine, a sphinx between the badge of the royal (Portuguese Order of the Tower and Sword) and the gold medal presented to Col. John Cameron of Fassifern by command of the Grand Signior, in testimony of that sovereign's high sense of his services in Egypt, and on a chief embattled a representation of the town of Aire in France, all proper (Cameron of Fassifern).
(Right) Representation of two coats of arms quartered.



Marks of bastardy.
These are common marks of illegitimacy but do not invariably have that meaning. (Left) The arms of the duke of St. Albans debruised by a baton sinister, in this case charged with three roses. (Right, top) The bordure wavy (or a bordure wavy sable). (Centre) The bordure compony (vert a bordure compony argent and gules). (Bottom) The baton sinister (purpure a baton sinister argent).
Drawing by Wm. A. Norman

of arms of dominion is to be found in the numerous members of the Commonwealth. Canada, for instance, has arms for the sovereign of Canada as authorized by George V in 1921, and the 12 provinces of Canada have similar arms approved by the sovereign. Much the same is true of the former French colonies, though there was no sovereign to grant the arms. The arms of political units are used throughout the Western world. The cities and boroughs of the United Kingdom, for example, have their heraldry, as do the U.S. states.

The blazoning of these nonfamilial arms is conducted on the same principles as for family arms, except that the explanation of the charges is usually forthcoming.

Drawing by Wm. A. Norman



The arms of Canada, derived from the royal arms. The three maple leaves in the base, originally green, were altered to red to conform to the new national flag adopted in 1964. French traditions and ties are represented by the gold fleurs-de-lis and the white lilies in the compartment.

Historical development of heraldry

The exact date or place of origin of the heraldic system in western Europe in the 12th century is not known; neither are the precise reasons for its introduction. But limits can be drawn to indicate generally when heraldry began, and probable reasons for its emergence can be deduced. The Bayeux Tapestry, produced in the last quarter of the 11th century in Bayeux, France, is a pictorial record that shows conditions in everyday life and in warfare at the time of the Norman Conquest of England. The English and Norman soldiers are armed alike. None of the Englishmen has a design of any kind on his shield or armour. In a few scenes, Normans or Frenchmen have designs on their shields that have a rough heraldic resemblance. In scene VIII of the tapestry, four of the followers of Guy, count of Ponthieu, have shields with these devices: some kind of creature holding what appears to be a fish in its mouth, a rough design emerging from the left side of the shield, a cross, and an animal resembling a sheep. In scene XII the messenger of William the Conqueror bears a winged creature on his shield, and this reappears in scenes XIII and XV. In scene XVIII a cross or a variant of it is seen on a shield, but this was probably a boss (metal protuberance) to strengthen it. Scene LXXV has a Norman knight with a design of a birdlike creature on the shield, but generally the Normans shields have only bosses in the middle.

Bayeux Tapestry and its record of shields

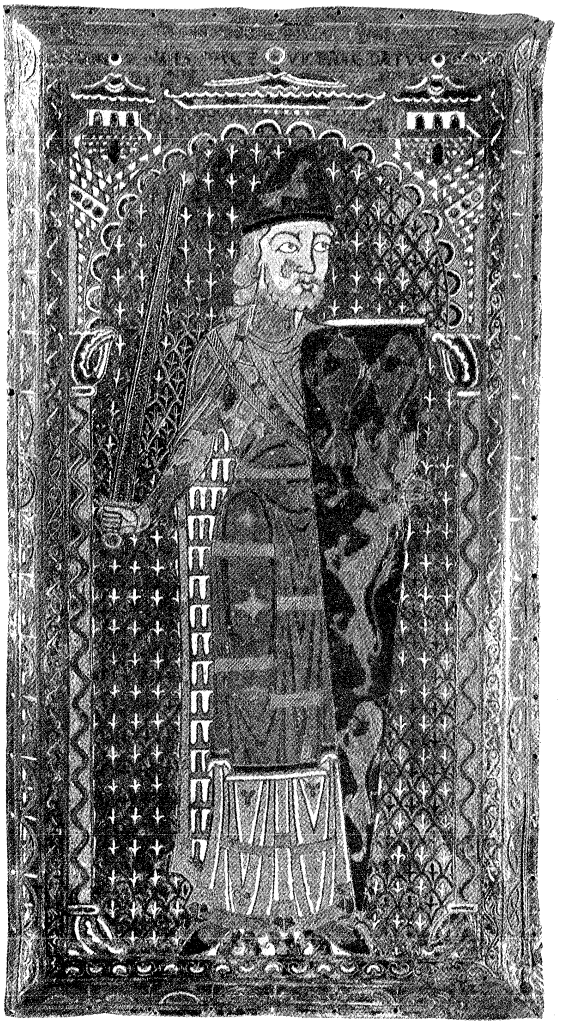
In 1066, at the time of the Conquest, heraldry clearly did not exist. The most that can be said is that possibly some of the rudiments out of which it emerged were present. Even at the time of the First Crusade (1095–99), there is evidence that heraldry was not yet in use. The *Alexiad,* the history of the Emperor Alexius (reigned 1081–1118) written by his daughter, the princess Anna Comnena, contains a vivid description of the Frankish barbarians—as the Crusaders appeared in the eyes of the civilized Byzantines. The Princess gave a careful account of the Crusaders' armour. She said that Alexius exhorted his archers to shoot at the Franks' horses rather than their riders, whose armour rendered them almost invulnerable. "For the Frankish weapon of defence is this coat of mail, ring plaited into ring, and the iron fabric is such excellent iron that it repels arrows and keeps the wearer's skin unhurt. An additional weapon of defence is a shield which is not round, but a long shield, very broad at the top and running out to a point, hollowed out slightly inside, but externally smooth and gleaming with a brilliant boss of molten brass." The shields in her description, therefore, are similar to those depicted in the Bayeux Tapestry. Thus, all the pictures from later times of King Alfred the Great, William the Conqueror, or Charlemagne bearing coat armour can be disregarded as anachronisms.

EARLY ROOTS OF HERALDRY

Not until a generation after the First Crusade does unmistakable evidence of heraldic designs appear. The earliest evidence is in an enamel (Musée de Tessé, Le Mans, France) made not later than 1151 showing Geoffrey, count of Anjou, bearing a shield azure with possibly four rampant golden lions (the exact number is not discernible because of the position in which the shield was depicted). The count was the son-in-law of King Henry I of England (reigned 1100–35), and, according to a chronicle, Henry in knighting Geoffrey bestowed upon him a shield that bore painted lions. In addition, from 1136, heraldic devices appear on seals. It is possible that the insignia were used first on seals and later on warriors' shields. Simultaneously, body armour was becoming all-enveloping, and some means of distinguishing men in full armour became necessary. The heavy barrel-type helmet closed in the wearer's face except for the opening of his visor, and a mixture of plate and ring mail enclosed the whole body. Another factor was the Crusades, in which men from different lands had to be distinguished from one another.

The first unmistakable evidence of heraldic designs

Within a few years heraldry was found throughout all of western Christendom. The first English king to bear arms was the crusader Richard I the Lion-Heart (1157–99). The three gold leopards or lions of England have been used by every dynasty since his time.



Plaque from the tomb of Geoffrey Plantagenet, count of Anjou, enamel, Limoges school, *c.* 1151–60. The stylized pattern of blue and white lining the figure's cloak represents a series of squirrel skins, called vair, frequently mentioned in blazons. In the Musée de Tessé, Le Mans, France.

Giraudon—Art Resource/EB Inc.

Seals. The earliest body of evidence of heraldic insignia is found in seals, large numbers of which have been preserved in England, France, and Germany, with fewer surviving in Spain and Italy. For the first century of heraldry, seals supply the bulk of information. It is from seals that the rise and development of the English royal arms can be traced. Seals from the first years of Richard I's reign show the design of a lion rampant to the left side. Some scholars think that two lions were used, since only half of the shield can be seen. Seals from the end of Richard's reign bear the three lions or leopards that have been used by all subsequent English sovereigns. The adoption of the same coat of arms by subsequent dynasties is also found in the royal arms of Sweden and of Denmark; but unlike the English, those royal families place their family arms on an inescutcheon in the middle of the shield.

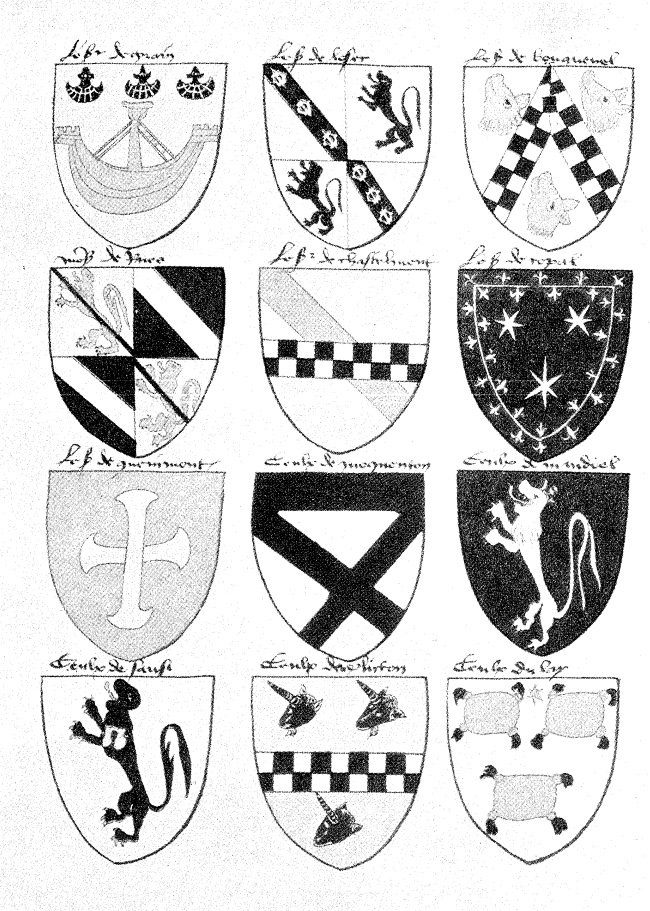
Roll of arms. Next to seals as evidence of heraldic usage come the rolls of arms, which in England date from about 1250. These are lists of arms, often with pictures drawn ("tricked") on the rolls, of persons who were present on a particular occasion, such as at a tournament or on a military expedition. England and Belgium (Flanders) are rich in the rolls of arms. France, Spain, and Scotland have fewer surviving examples. In place of the rolls, collections of painted books of arms have been preserved in Germany. A notable roll is the *Armorial de Berry,* dating about 1445, the work of a French herald, Gilles le Bouvier, who travelled widely and recorded arms borne in France, England, Scotland, Germany, Italy, and other European countries.

The *Armorial*

Records in stone and glass. Another very important source of information is to be found in representations on stone, wood, glass, and in books and engravings. Over the gateway of Bodiam Castle in Sussex can be seen the arms cut in stone of three owners of the castle, the families of Bodiam (who took their name from the place), Wardedieux, and Dalyngrygge. Of such arms nothing would be known without these centuries-old memorials. In Rome, many examples occur of the arms of various popes in their palaces and other buildings, for instance the bees of the Barbarini pope Urban VIII in the Palazzo Venezia. Heraldic glass is usually much more recent in origin but of immense value in supplying information as it is always in colour, while other memorials often are not. Very few churches of any great age in western Europe are without armorial illustration. Switzerland in particular has splendid memorials in stained glass; for example, the Dom, or main Protestant church in Berne, has windows that are aflame with glorious heraldic colours. Sweden has a fine collection of coloured plaques of arms in the House of Nobles in Stockholm; in the Frederiksborg Slot (castle) at Hillerød, Denmark, may be seen the shields of the Knights of the Order of the Elephant, in which can be read the history of heraldry over several centuries.

Church brasses. Brasses in churches are an important source of heraldic information. It was formerly the custom to put a brass tablet over the grave slab, and on this would be shown a figure of the deceased with his armorial bearings. Many fine examples of this are found in old English churches. A very fine collection of floor brasses is in the small church of Stopham in Sussex, which has been the memorial place of the local Barttellot family for many centuries. Also found in churches are hatchments, heraldic paintings on wood that were made for deceased persons and hung over their house doors, being later set up in the local church where they have often been preserved.

By courtesy of the Society of Antiquaries, London



Page from the *Armorial de Berry,* by Gilles le Bouvier *c.* 1445, showing the simplicity of the early coats. In the collection of the Society of Antiquaries, London.



Sir Nicholas Hawberk, rubbing from his tomb brass, Cobham Church, Kent, 1407. Hawberk's arms appear alone on the dexter and impaled with the arms of his wife on the sinister.
By courtesy of Robert Lind

As for written material—such as official enactments, grants of arms, and books—nothing is dated earlier than the 14th century.

GROWTH OF HERALDRY AFTER THE 13TH CENTURY

The initial meaning of the term herald is uncertain. Some authorities derive the word from two German words—*Heer,* "a host," and *Held,* "a champion"—not a very obvious etymology. It is clear that heralds were in existence from the 13th century, if not earlier. In their beginnings they were more or less menials, ranked with the jugglers and the minstrels. First used as messengers who wore livery, they later were used in the jousts and tourneys to announce the contenders. To identify the knights, they had to know the arms on the shields, and from this grew their knowledge and skill in heraldry.

Messengers and aides at jousts and tourneys

Heraldic colleges and offices. From this lowly origin have come the colourful figures of the English College of Arms, who now alone, save for the Scottish heralds, possess a high position in the modern world. The Lord Lyon, the head of the Scottish heralds, derives his office from a much higher source than do the heralds in other parts of Europe. The Sennachie, or official bard of the Scots' king, was the record keeper of the old Celtic kingdom of Scotland, and from the Sennachie is derived the Lord Lyon, a great officer of state in Scotland.

The older statements found in many books that the medieval heralds were either identical, or in some way connected, with the old Greek *kēryx* or Latin *fetialis* need only be stated to be dismissed. Since ancient times men have been found who, because their persons were accepted as sacred, were able to carry messages and other communications between nations either hostile or strange to one another. These ambassadors bore several names before the development of a diplomatic corps. In the earlier Middle Ages, for instance, churchmen, monks, or priests were used for this type of service. When William I the Conqueror sent a messenger to Harold II of England, it was a monk who carried William's denunciation of Harold. Heralds were not then in existence.

As they ascended the social scale, heralds began to serve as ambassadors between the different courts, a function that was still theirs in the first half of the 17th century. In 1627, for example, Sir Henry St. George was joined in a commission with Lord Spencer and Peter Young to present the insignia of the Order of the Garter to Gustavus II Adolphus, king of Sweden, who then knighted Sir Henry and granted him an augmentation to his arms showing the royal arms of Sweden.

At first every great noble had his herald, and the royal heralds were distinguished from the others by the greater importance of their masters. Gradually, it came about that a king would form his heralds into a college or corporation. The King of France did so in 1407; it was not until 1484 that the King of England followed by establishing the College of Arms, which has been housed for 300 years in the same building in London. The English College, sometimes called the Heralds' College, has outlived all similar elaborate establishments in Europe, except that in Scotland. Outside Great Britain, heraldic offices are found in the 20th century in Sweden, Denmark, the Republic of Ireland, and Spain.

The English College

The English College is under the control of the earl marshal, an office that has been hereditary in the family of the duke of Norfolk since the 1660s. The holder of the dukedom is always earl marshal. Under him are 13 officers of arms; three kings of arms (Garter, Norroy and Ulster, and Clarenceux); six heralds (Windsor, Richmond, York, Lancaster, Chester, and Somerset); and four pursuivants (Rouge Dragon, Rouge Croix, Bluemantle, and Portcullis). These medieval names are derived from sources connected with royalty, titles, badges, or orders of knighthood. Pursuivants are "followers," or junior heralds. In Scotland the Lord Lyon is the head of the heraldic officers, of whom there are three heralds (Albany, Marchmont, and Rothesay) and four pursuivants (Carrick, Kintyre, Unicorn, and Ormond). In England and Scotland the officers are not civil servants but members of the sovereign's household. In both countries there are also heralds extraordinary, who are appointed at times for special reasons or functions.

Records and grants. In England an important development came with the Heraldic Visitations. From 1530 in the reign of Henry VIII to 1686 in the reign of James II, commissions were issued by the sovereign to the heralds directing them to proceed to a county in England or Wales and to inspect the arms in use there. The records of the Visitations have been preserved and constitute a valuable body of genealogy as well as of heraldry. From the period of the Visitations the heralds built up huge collections of family history and began to record pedigrees.

Heraldic Visitations from 1530 to 1686

From about a half century before the foundation of the College of Arms, the English heralds are found to be issuing grants of arms on behalf of the sovereign. This is some 300 years after the first appearance of heraldry, which obviously much antedated not only royal colleges or corporations of heralds but even the existence of heralds themselves. From this evidence, it seems clear that in the early days of heraldry men assumed arms to suit themselves without reference to any authority. A very simple coat of arms would not be difficult to invent. That three unrelated persons from three different counties could bear these same arms is not only not surprising but proof that the arms were self-chosen. When disputes over ownership among the three came up, the matter was referred to the king. His judgment was final, but it is noteworthy that one of the defeated, compelled to give up his arms, then consoled himself with a new coat that was also self-derived and self-assumed. Unquestionably, the great majority of ancient coats of arms, borne before 1500, were never granted but were taken by the owners.



Detail of the "Grant of arms to the Worshipful Company of Tallow Chandlers," London, 1456. The herald in the illuminated initial wears a tabard that bears the king's arms front, back, and repeated on the sleeves, and a coronet as a mark of royal authority. In the collection of the Worshipful Company of Tallow Chandlers, London.

By courtesy of the Court of the Worshipful Company of Tallow Chandlers, London

Writers on heraldry. The earliest writing on heraldry extant is a short treatise by Bartolo da Sassoferrato, whose *Tractatus de insigniis et armis* ("Treatise on Insignia and Arms") was published about 1356. In his small book Bartolo describes the various categories of arms bearing and how they have been assumed. He refers specifically to arms granted by a prince and gives reasons for their value but asks why one man may not bear arms identical with those of another.

In 1355 Bartolo had been sent to Pisa from Perugia as an envoy to the Holy Roman emperor, Charles IV, from whom he received many privileges, including a grant of arms, which were the same as those of the Emperor as king of Bohemia but with changed tincture: "or a lion rampant with two tails gules." An American scholar, L.M. Mladen, remarked of this grant and others made by Charles IV at the same time: "Charles was in all probability the first ruler ever to grant arms. To my knowledge, no earlier occurrence has been found."

The first English heraldic writer was John of Guildford, or Johannes de Bado Aureo, whose *Tractatus de armis* ("Treatise on Arms") was produced about 1394. Then came a Welsh treatise, the *Llyfr arfau* ("Book of Arms"). Nicholas Upton, a canon of Salisbury Cathedral, about 1440 wrote *De studio militari* ("On Military Studies"). John of Guildford's treatise was printed in 1654 with Upton's work and the *Aspilogia* of Sir Henry Spelman by Sir Edward Bysshe, Garter king of arms, who edited and annotated all three works. The whole was in Latin; no complete English version of Upton's book has been published.

The trend to the nonsensical after the 15th century

These books are by authorities who were concerned with the realities of heraldry in their own day. A tendency away from actuality and toward the fanciful and absurd manifested itself from the end of the 15th century. Some of these farfetched conceits showed themselves in *The Boke of St. Albans* (1486). Yet by comparison with the vast mass of nonsense contained in the folios of the 16th century, those conceits were reasonable. The works of Sir John Ferne, *Blazon of Gentrie* (1586), Gerard Legh, *The Accedens of Armorie* (1562), and John Guillim, *A Display of Heraldrie* (1610) not only perpetuate the nonsensical natural history of olden days but are largely responsible for the belief that heraldic charges have a definite symbolic meaning and that they were granted as the reward of valorous deeds.

Continental versus British heraldry. Much greater significance was attached in former times to heraldic insignia, though the attitude varied from country to country. Heraldry has become more widespread than at any other time, but as a sign of rank it has hardly any remaining value.

A distinction can be made between the Continent and Great Britain regarding medieval and later heraldry. The doctrine of the Seize Quartiers (16 quarterings) prevailed over most of the Continent but not in Britain. This theory required that, in order for a person to claim nobility, all of his 16 ancestors (16 great-great-grandparents) should have been entitled to bear arms. This is known as the "Proof of the Seize Quartiers" and was the reason why Frederick II the Great of Prussia, though professing the views of the Enlightenment in the Age of Reason, diligently scrutinized his courtiers' quarterings. The theory is based on the rigidity of a noble caste that married only with its own kind. On the Continent, every member of a noble family is noble; hence the enormous numbers of titles. Similarly, the continental royalty tended to marry only with other royal families. As a result both royal and noble families formed a class apart from the bulk of the people.

Seize Quartiers

Continental heraldic insignia, therefore, from their origins until the late 18th century, provided symbols to indicate a higher caste and, in fact, were signs of nobility. Yet, strangely, in several countries heraldry was in wide general use as a means of identification, serving in the same way as a surname. In France, for example, it is abundantly clear that from the 13th century, not only the bourgeoisie of the towns but also the peasants bore heraldic arms. The usage had percolated down from the noble class. The earliest example of the use of arms by a peasant is that of Jaquier le Brebiet in 1369, whose arms show a punning allusion to his name (*brebis,* "sheep"—three sheep held by a girl). In other European lands—Hungary and the Low Countries—burgher or peasant arms were also found, but neither in these lands nor in France were the possessors regarded as noble.

Peasant and bourgeois arms

In France the regulations of arms followed a quite different course from those in England. Although King Charles VI had in 1407 led the way in creating a college of arms, his heralds lost influence over the next two centuries. They had no power, unlike their Scottish and English counterparts, to grant arms and they gradually faded into insignificance. To overcome the loss, Louis XIII in 1615 appointed a *juge géneral d'armes* ("general judge of arms"), an official whose powers resembled those of the Lord Lyon. The French royal government showed itself very broad-minded on the possession of arms. A decree of Louis XIV in 1696, designed to raise money, ordered all persons who bore arms to register them. Even those who were not part of the arms-bearing population were forced to buy arms. Later, in 1760, an ordinance was framed by which the lesser townsfolk, artisans, and peasants were to be excluded from the use of arms—after these classes had used them for 400 years. But the Parlement of Paris refused to allow the ordinance to be implemented. Later, during the French Revolution (1789), arms were suppressed as signs of feudalism.

The view of arms held in England was and is quite different. No such thing as a noble caste has ever existed in England. Only the reigning peer and his wife are regarded as noble; the rest of the family are commoners, and only a few of them bear what are called courtesy titles. Moreover, except for the Hanoverian (1714–1837) and Victorian (1837–1901) epochs, the royal house has not necessarily married royalty but much more often the nobility, a practice to which it has returned in the present century. As a result, the Continental doctrine of Seize Quartiers does

not apply in England. It cannot, since noble and nonnoble are so mixed. Nor are there any such things as nonnoble arms. All arms are on the same basis; all are signs of gentility—nobility, in fact. Arms then have long had a high social significance in England; those who possess them have social prestige. The situation is somewhat different in Scotland where arms are bound up with clanship and family solidarity. To the Scotsman arms are not so much a matter of social status as of family or clan, and he proves his right to them by process of law in Lyon Court.

A matter of clan and family in Scotland

20TH-CENTURY HERALDRY

France and Italy. Without a monarchy, heraldry can still flourish, but it does not normally do so. The French Revolution abolished arms, which returned with the monarchy, and now, although France is a republic, a person assuming arms to which he is not entitled may be prosecuted. In the same way, it is not permissible to assume a name of a great family. In England, however, the assumption of the name Windsor (the name of the British royal family) is possible for anyone. There is, however, no recognized heraldic authority in France, nor for that matter in other European countries that have abolished their monarchies. Most such republican states have associations that seek to maintain heraldic standards. Thus in Italy the Collegio Araldico (Heraldic College) consists of experts whose main object is to promote heraldic and genealogical studies. An association of nobles was formed as the National Heraldic Council of the Italian nobility, under the authority of former king Umberto II; it tries to regulate the use of arms and titles.

Many changes in heraldry have taken place since 1945, as it is not a static subject.

Communist countries. In Communist European countries, the study of genealogy and heraldry has been generally suppressed. Since about 1956 it has been possible to obtain much more statistical information from the U.S.S.R. than formerly, but no heraldic data are supplied. Much the same is true of the other Communist countries in Europe, except Czechoslovakia. In Communist lands, heraldry is viewed as it was by the French revolutionaries, as part of the feudal past, although there is reason to believe that in these countries heraldic archives are carefully preserved.

Ireland. A different development occurred in two of the republics that have emerged from the British Empire—Ireland and South Africa—both of which have set up their own heraldic offices. As early as 1382, there was an Ireland King of Arms who was responsible for all matters armorial in that country. The last holder of the office died in 1487, and in 1553 Edward VI created a new armorial king under the title of Ulster, to control bearings throughout Ireland. His place of business was in Dublin Castle. When the Irish Free State, or Eire (now the Republic of Ireland), was established in 1921–22, the Ulster Office was reserved as an appointment of the British crown with the then-current Ulster to hold office for life. After his death in 1940, an arrangement was made between the British and Irish governments by which the heraldic office in Dublin Castle with its records was taken over by the Irish authorities. Photostat copies were made of the records and sent to the College of Arms, London. The Irish government appointed a Chief Herald of Ireland, and the Ulster Office became known as the Genealogical Office. A civil servant was then appointed as Chief Herald of Ireland. The office of Ulster King has now been united with that of Norroy King in the College of Arms, London. The Irish Herald carries out the duties formerly performed by Ulster in the 26 counties of the Republic of Ireland; Norroy and Ulster has jurisdiction over the six counties of Northern Ireland (Ulster).

Chief Herald of Ireland

South Africa. In South Africa an act was passed in 1962 under which was established a Bureau of Heraldry and a Heraldry Council for the grants, registration, and protection of coats of arms, badges, and other emblems. A state herald is appointed as head of the Bureau of Heraldry. The Heraldry Council consists of the state herald and at least seven other members appointed by the government minister responsible.

The United States. There has been a remarkable evolution of heraldry in the United States. Ever since the American Revolution, the use of arms, especially of arms of English families with whom the users were related or whose surname they bore, has continued. The English College of Arms claims heraldic jurisdiction over persons of English and Welsh descent (Wales has been reckoned with England in this and all other administrative matters since the union of England and Wales, 1542). The Lord Lyon of Scotland claims jurisdiction likewise over persons of Scottish descent throughout the world. In addition, the English College at one time claimed a worldwide imperial jurisdiction over anyone who could be brought within the definition of British subject. Under this jurisdiction even the Indian princes were occasionally granted arms by the College of Arms, although they were not British subjects but independent rulers who had entered into treaty relations with the British crown. Many Americans have been granted arms by the college by virtue of their descent from English or Welsh forebears, or by the Lord Lyon if they are of Scottish descent. Irish Americans often were granted arms from Dublin, either from the Ulster King or his successor. Americans of Northern Irish descent have been granted arms by Norroy and Ulster. In addition, there are several states of the United States that were formerly Spanish territory, and the Spanish Kings of Arms, the equivalent of the English and Scottish heralds, exercised a heraldic authority over persons of Spanish descent in the old Spanish Empire. By extension, they have recently granted arms to Americans who are resident in those formerly Spanish states but who are not of Spanish descent.

Heraldry in states from the old Spanish Empire

To these classes of arms obtained by Americans from overseas must be added such instances as the award of arms to Pres. Dwight D. Eisenhower in Denmark. Also, Americans of French, German, Italian, Polish, and other European origins have inherited from their immigrant ancestors arms once granted or recorded by heraldic authorities no longer in existence. All these classes of arms share one feature whatever their origin: they are hereditary honours granted to American citizens by other countries. As they do not carry titles, they do not contravene the principle of the American Constitution on this subject. An American who receives a knighthood of some foreign state possesses only an honorary knighthood; he is not "Sir." But for the citizen of an independent sovereign power to approach and receive from another power a hereditary honour has seemed to many Americans an undesirable procedure. There have been and still are thousands of assumptions of arms by Americans on consideration of mere mail-order salesmanship; arms in these transactions are, at the best, supposed to be those "of the name" (that is, belonging to a name rather than a person), a view for which no justification exists.

Endeavours have been made to set up American authorities who would not only record but also grant arms. The New England Historic Genealogical Society of Boston appointed a Committee on Heraldry that since 1928 has issued rolls of arms, in which have been entered the names and arms of those who have submitted their claims to its judgment. The use of this method of issuing or publicizing arms recalls the usage previously referred to, which has not been practiced in Europe for more than 400 years. In the introduction to the second roll (1932) it is stated: "There is certainly no legal reason, perhaps no reason at all, why an American gentleman should not assume *in more majorum* any new coat that pleases his fancy, but he should not assume an old coat, for if he does, he is very likely denying his own forefathers and he surely is affirming what he has no sufficient reason to be true." Not only British-derived arms but also continental European arms have been registered. In addition, the committee has assisted inquirers in devising new coats of arms, not only for schools, colleges, and other institutions, but also for individuals. In the introduction to the first roll a very reasonable view toward heraldry was expressed: "Taking into consideration the early history of coat armour there seems to be no reason in this country at least, why anyone provided he observes the simple rules of blazon and does not appropriate the arms of another, may not assume and use

any coat he desires." The American College of Heraldry and Arms, Inc., was established in the state of Maryland in 1966. It has two divisions: The American College of Arms, which is concerned with the arms of individuals, their registration, and, more importantly, the granting of arms; and the College of Arms of the United States, which deals with such items as arms, crests, and standards for corporate concerns. Arms were granted to Pres. Lyndon B. Johnson during his term of office, to Pres. Richard M. Nixon, and to Vice Pres. Spiro T. Agnew.

Revival of England's ancient Court of Chivalry

England. Heraldic development has also occurred in England, where in 1954 the ancient Court of Chivalry was revived. This was once the court of the lord high constable and the earl marshal, and it dealt with matters relating to knights and gentlemen. Although it was concerned also with matters of military discipline, it was not the forerunner of the modern court-martial in the armed forces. The court gradually declined in the 17th and 18th centuries and had not sat from 1735 until its revival.

The office of lord high constable has long ceased to be hereditary or of permanent status in England. During coronations, however, a constable is appointed for the occasion. Therefore, in the revived court that sat in 1954 to deal with a test case, *Manchester Corporation* v. *Manchester Palace of Varieties,* the earl marshal (the duke of Norfolk) presided with a surrogate, who was the lord chief justice of England appearing in his capacity as a doctor of civil law. As a result of the sitting, the jurisdiction of the court was confirmed, and the City of London recorded its crest and supporters.

The unorganized condition of heraldry in many European countries has spurred private attempts to bring some order into the field. The movement known as the International Congresses of Heraldry and Genealogy began in 1928 with a meeting in Barcelona, Spain. A second Congress was held in Rome and Naples in 1953, and from that time regular meetings occurred at two- or three-year intervals. Out of these was established an international organization, El Instituto Internacional de Genealogia y Heraldica (The International Institute of Genealogy and Heraldry), with its headquarters in Madrid.

USES OF HERALDRY FOR STUDY AND VERIFICATION

The uses of heraldry, apart from its general significance in providing distinguished symbols, are considerable. Heraldry illustrates much history and literature that is otherwise obscure. Heraldry on buildings, in manuscripts, and in paintings is of immense value for purposes of identification. It serves to link one person with another, to connect families, and to disclose origins of states and of institutions. With the use of heraldry in connection with genealogy, from which it cannot easily be separated, much that is difficult to follow becomes easy to interpret. In every building that contains armorial engravings or other pictures of arms, there is a concise history of the place, which gives a knowledgeable onlooker a useful clue to the background and details of the building.

In law the place of heraldry is sometimes very precise, as in modern Scotland; in England, on the other hand, it is more involved. To understand the latter is to gain an insight into the development of English law. Similarity of arms does not always indicate identity of family; it may merely denote assumption of devices. According to the English laws of inheritance, not only an estate but also a surname and arms can pass by eventual succession to individuals unconnected by blood with the original owner. Thus in the families of Lytton and Carew, and in branches of Trelawny, instances occur in which possession of the name and arms is contrary to blood relationship. Sometimes, however, identity of a name with a puzzling discrepancy in the arms can only be explained by a study of the family's heraldry.

BIBLIOGRAPHY. There are many guides to heraldry; the majority appear to be derived from an old original. Among the most useful for the beginner are: A.C. FOX-DAVIES, *Heraldry Explained* (1906, reprinted 1971), a small, well-illustrated book written by the clearest expositor of heraldry; CHARLES MACKINNON, *The Observer's Book of Heraldry* (1966), a very useful and clearly written work; and SIR ANTHONY R. WAGNER, *Heraldry in England* (1946, reprinted 1953), a brief account but with 15 plates in colour; LESLIE G. PINE, *Teach Yourself Heraldry and Genealogy,* 2nd ed. (1970), designed for the beginner and including a glossary of terms, and *The Genealogist's Encyclopaedia* (1969, reprinted 1977); and JAMES R. PLANCHÉ, *The Pursuivant of Arms or Heraldry Founded upon Facts* (1852, reprinted 1973).

More specialized works include A.C. FOX-DAVIES, *A Complete Guide to Heraldry* (1909, reprinted 1978), and *Armorial Families: A Directory of Gentlemen of Coat Armour,* 7th ed., 2 vol. (1929–30, reprinted 1970). GERARD J. BRAULT, *Early Blazon: Heraldic Terminology in the Twelfth and Thirteenth Centuries, with Special Reference to Arthurian Literature* (1972), illustrates the development of the language of blazon. The writings of OSWALD BARRON should be read as a correction to some of Fox-Davies' theories on heraldic history. Some of his best work may be found in the 12 volumes of *The Ancestor,* which he edited in 1902–05. Two volumes now reprinted in one constitute *A Treatise on Heraldry: British and Foreign* by JOHN WOODWARD and GEORGE BURNETT (1892, reprinted 1969), which contains some inaccuracies. SIR THOMAS INNES, *Scots Heraldry,* new ed. (1978), explains much in heraldic practice that may not otherwise be clear. Another good view of heraldry in the international sense is ROBERT GAYRE OF GAYRE AND NIGG, *The Nature of Arms* (1961). For information on Irish heraldry, see SIR CHRISTOPHER and ADRIAN LYNCH-ROBINSON, *Intelligible Heraldry* (1948, reprinted 1967). *Boutell's Heraldry* has been edited many times since its first appearance in the 19th century. The editions of V. WHEELER-HOLOHAN (1931), C.W. SCOTT-GILES (1958), and JOHN P. BROOKE-LITTLE (1978) are all very useful, as are JOHN P. BROOKE-LITTLE, *An Heraldic Alphabet* (1973); and OTTFRIED NEUBECKER and JOHN P. BROOKE-LITTLE, *Heraldry: Sources, Symbols, and Meaning* (1977). JULIAN FRANKLYN, *Shield and Crest,* 3rd ed. (1971), gives much otherwise not easily obtained information.

Little continental work on heraldry has been translated into English. Perhaps the best course is to consult periodical publications like *The Augustan,* with articles from a wide variety of sources. Several non-English-language encyclopaedias, such as the *Enciclopedia Italiana* (1929–37), have excellent articles on heraldry. Many non-English writers transcend their national boundaries in writing on the subject. RÉMI MATHIEU in *Le Système héraldique français* (1946) writes of French heraldry and helps to clarify the whole subject. On Spanish heraldry, see JOSÉ ASENSIO Y TORRES, *Tratado de heráldica y blasón,* 3rd ed. rev. (1854, reprinted 1929); and LUCAS DE PALACIO, *De Genealogia y heraldica* (1946), the latter author being especially interested in the connection between totemism and heraldry; the Japanese *mon* is exhaustively dealt with by CARROLL PARISH in *The Augustan,* vol. 11, no. 1 (1968).

Other works that will help the student after he has acquired a sound knowledge of this subject are: SIR ANTHONY R. WAGNER, *Heralds and Heraldry in the Middle Ages,* 2nd ed. (1956, reprinted 1960), and *Historic Heraldry of Britain* (1939, reprinted 1972); and LESLIE G. PINE, *The Story of Heraldry* (1952, reprinted 1967), an account of much controversial matter, including the present English heraldic position in law; GEORGE D. SQUIBB, *The High Court of Chivalry: A Study of the Civil Law in England* (1959); C. PAMA, *Lions and Virgins* (1965), which discusses the history of arms in South Africa; and LESLIE G. PINE, *International Heraldry* (1970), and *American Origins* (1960, reprinted 1980), concerned with heraldry throughout the world.

Reference works in English are easily available. SIR JOHN B. BURKE, *The General Armory of England, Scotland, Ireland, and Wales* (1883, reprinted 1976), gives the description under alphabetical order of surnames of thousands of coats of arms. JOHN WOODY PAPWORTH, *An Alphabetical Dictionary of Coats of Arms Belonging to Families in Great Britain and Ireland* (1858–74, reprinted 1977), the counterpart of Burke's *General Armory,* enables the seeker to trace a coat of arms without knowing the owner's name. Both books contain inaccuracies. JAMES PARKER, *A Glossary of Terms Used in Heraldry,* new ed. (1970), with 1,000 illustrations, is a helpful book, as is LESLIE G. PINE, *A Dictionary of Mottoes* (1983). Regular editions of JOHN DEBRETT, *Peerage, Baronetage, Knightage and Companionage* (since 1713); and JOHN B. BURKE, *A Genealogical and Heraldic History of the Peerage, Baronetage, and Knightage* (since 1826), and *Genealogical and Heraldic History of the Landed Gentry* (since 1837), abound in illustrations and descriptions of arms. The role of the herald in history and in the contemporary art and science of heraldry is explored in RODNEY DENNYS, *The Heraldic Imagination* (1976), and *Heraldry and the Herald* (1982).

(L.G.P.)

Hinduism

The term Hinduism refers to the civilization of the Hindus (originally, the inhabitants of the land of the Indus River). Introduced in about 1830 by British writers, it properly denotes the Indian civilization of approximately the last 2,000 years, which evolved from Vedism, the religion of the Indo-European peoples who settled in India in the last centuries of the 2nd millennium BC.

Because it integrates a variety of elements, Hinduism constitutes a complex but largely continuous whole and has religious, social, economic, literary, and artistic aspects. As a religion, Hinduism is a composite of diverse doctrines, cults, and ways of life.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* section 823, and the *Index.*

This article is divided into the following sections:

General nature and characteristic features 519
Common characteristics of Hindu belief 519
Three margas: paths to salvation 520
The history of Hinduism 521
Sources of Hinduism 521
The prehistoric period (3rd and 2nd millennia BC) 522
The Vedic period (2nd millennium–7th century BC) 522
Challenges to Brahmanism (7th–2nd century BC) 522
Early Hinduism (2nd century BC–4th century AD) 523
The rise of devotional Hinduism (4th–11th century) 524
Hinduism under Islām (11th–19th century) 525
The modern period (19th–20th century) 527
Sacred texts 529
Vedas 529
Sutras, Shastras, and Smritis 533
Epics and Purāṇas 534
Vaishnavism and Śaivism 537
Philosophical texts 540
Tantrism 542
Vernacular literatures 545
Folk Hinduism 547
Rituals, social practices, and institutions 549
Sacrifice and worship 549
Sacred times and places 551
Ritual and social status 552
Cultural expressions: visual arts, theatre, and dance 554
Types of symbols 554
The arts 554
The place of Hinduism in world religions 555
Hinduism and other religions of Indian origin 555
Hinduism and Islām 556
Hinduism and Christianity 556
Bibliography 557

General nature and characteristic features

The spectrum that ranges from the level of popular Hindu belief to that of elaborate ritual technique and philosophical speculation is very broad and is attended by many stages of transition and varieties of coexistence. Magic rites, animal worship, and belief in demons are often combined with the worship of more or less personal gods or with mysticism, asceticism, and abstract and profound theological systems or esoteric doctrines. The worship of local deities does not exclude the belief in pan-Indian higher gods or even in a single high God. Such local deities are also frequently looked upon as manifestations of a high God.

Religious pluralism

In principle, Hinduism incorporates all forms of belief and worship without necessitating the selection or elimination of any. It is axiomatic that no religious idea in India ever dies or is superseded—it is merely combined with the new ideas that arise in response to it. Hindus are inclined to revere the divine in every manifestation, whatever it may be, and are doctrinally tolerant, allowing others—including both Hindus and non-Hindus—whatever beliefs suit them best. A Hindu may embrace a non-Hindu religion without ceasing to be a Hindu, and because Hindus are disposed to think synthetically and to regard other forms of worship, strange gods, and divergent doctrines as inadequate rather than wrong or objectionable, they tend to believe that the highest divine powers complement one another. Few religious ideas are considered to be irreconcilable. The core of religion does not depend on the existence or nonexistence of God or on whether there is one god or many. Because religious truth is said to transcend all verbal definition, it is not conceived in dogmatic terms. Moreover, the tendency of Hindus to distinguish themselves from others on the basis of practice (orthopraxy) rather than doctrine (orthodoxy) further de-emphasizes doctrinal differences.

Hinduism is both a civilization and a congregation of religions; it has neither a beginning or founder, nor a central authority, hierarchy, or organization. Every attempt at a specific definition of Hinduism has proved unsatisfactory in one way or another, the more so because the finest scholars of Hinduism, including Hindus themselves, have emphasized different aspects of the whole.

COMMON CHARACTERISTICS OF HINDU BELIEF

Nevertheless, it is possible to discern among the myriad forms of Hinduism several common characteristics of belief and practice.

Authority of the Veda and the Brahman class. Perhaps the defining characteristic of Hindu belief is the recognition of the Veda, the most ancient body of religious literature, as an absolute authority revealing fundamental and unassailable truth. At the same time, however, its content has long been practically unknown to most Hindus, and it is seldom drawn upon for literal information or advice. Still, it is venerated from a distance by every traditional Hindu, and those Indians who reject its authority (such as Buddhists and Jains) are regarded as unfaithful to their tradition. The Veda is also regarded as the basis of all the later Shastraic texts used in Hindu doctrine and practice. Parts of the Veda are still quoted in essential Hindu rituals, and it is the source of many enduring patterns of Hindu thought.

Also characteristic of Hinduism is the belief in the power of the Brahmans, a priestly class possessing spiritual supremacy by birth. As special manifestations of religious power and as bearers and teachers of the Veda, Brahmans are considered to represent the ideal of ritual purity and social prestige.

Doctrine of atman-brahman. Hindus believe in an uncreated, eternal, infinite, transcendent, and all-embracing principle, which, "comprising in itself being and nonbeing," is the sole reality, the ultimate cause and foundation, source, and goal of all existence. This ultimate reality is called *brahman.* As the All, *brahman* causes the universe and all beings to emanate from itself, transforms itself into the universe, or assumes its appearance. *Brahman* is in all things and is the Self (atman) of all living beings. *Brahman* is the creator, preserver, or transformer

Identity of individual self and the ultimate reality

and reabsorber of everything. Although it is Being in itself, without attributes and qualities and hence impersonal, it may also be conceived of as a personal high God, usually as Vishnu (Viṣṇu) or Śiva. This fundamental belief in and the essentially religious search for ultimate reality—*i.e.,* the One that is the All—have continued almost unaltered for more than 30 centuries and have been the central focus of India's spiritual life.

Ahimsa: non-injury. A further characteristic of Hinduism is the ideal of ahimsa. Ahimsa, "non-injury" or the absence of the desire to harm, is regarded by Indian thinkers as one of the keystones of their ethics. Historically, ahimsa is unrelated to vegetarianism; in ancient India, killing people in war or in capital punishment and killing animals in Vedic sacrifices were acceptable to many people who for other reasons refrained from eating meat. However, the two movements, ahimsa and vegetarianism, reinforced one another through the common concept of the disinclination to kill and eat animals, and together they contributed to the growing importance of the protection and veneration of the cow, which gives food without having to be killed. Neither ahimsa nor vegetarianism ever found full acceptance. Even today, many Hindus eat beef, and nonviolence (as the ideal of ahimsa is often translated) has never been a notable characteristic of Hindu behaviour.

Doctrines of transmigration and karma. Hindus generally accept the doctrine of transmigration and rebirth and the complementary belief in karma, or previous acts as the factor that determines the condition into which a being, after a stay in heaven or hell, is reborn in one form or another. The whole process of rebirths is called samsara. Any earthly process is viewed as cyclic, and all worldly existence is subject to the cycle. Samsara has no beginning and, in most cases, no end; it is not a cycle of progress or a process of purification but a matter of perpetual attachment. Karma, acting like a clockwork that, while running down, always winds itself up, binds the atmans (selves) of beings to the world and compels them to go through an endless series of births and deaths. This belief is indissolubly connected with the traditional Indian views of society and earthly life, and any social interaction (particularly those involving sex or food) results in the mutual exchange of good and bad karma. It has given rise to the belief that any misfortune is the effect of karma, or one's own deeds, and to the conviction that the course of world history is conditioned by collective karma.

Samsara and moksha (liberation)

Such doctrines encourage the view that mundane life is not true existence and that human endeavour should be directed toward a permanent interruption of the mechanism of karma and transmigration—that is, toward final emancipation (moksha), toward escaping forever from the impermanence that is an inescapable feature of mundane existence. In this view the only goal is the one permanent and eternal principle: the One, God, *brahman,* which is totally opposite to any phenomenal existence. Anyone who has not fully realized that his being is identical with *brahman* is thus seen as deluded. The only possible solution consists in the realization that the kernel of human personality (atman) really is *brahman* and that it is their attachment to worldly objects that prevents people from reaching salvation and eternal peace. (Hindus sometimes use the largely Buddhist term nirvana to describe this state.)

Concepts of iṣṭadevatā and Trimūrti. Although those Hindus who particularly worship either Vishnu or Śiva generally consider one or the other as their "favourite god" (*iṣṭadevatā*) and as the Lord (Īśāna) and Brahman in its personal aspect, Vishnu is often regarded as a special manifestation of the preservative aspect of the Supreme and Śiva as that of the destructive function. Another deity, Brahmā, the creator, remains in the background as a demiurge. These three great figures (Brahmā, Vishnu, and Śiva) constitute the so-called Hindu Trinity (Trimūrti, "the One or Whole with Three Forms"). This conception attempts to synthesize and harmonize the conviction that the Supreme Power is singular with the plurality of gods in daily religious worship. Although the concept of the Trimūrti assigns a position of special importance to some great gods, it never has become a living element in the religion of the people. Moreover, Brahmā has had no major cult since ancient times, and many Hindus worship neither Śiva nor Vishnu but one or more of the innumerable other Hindu gods.

Ashramas: the four stages of life. In the West, the so-called life-negating aspects of Hinduism have often been overemphasized. The polarity of asceticism and sensuality, which assumed the form of a conflict between the aspiration to liberation and the heartfelt desire to have descendants and continue earthly life, manifested itself in Hindu social life as the tension between the different goals and stages of life. The relative value of an active life and the performance of meritorious works (*pravṛtti*) as opposed to the renunciation of all worldly interests and activity (*nivṛtti*) was a much-debated issue. While one-sided religious and philosophical works, such as the Upanishads, placed emphasis on renunciation, the dharma texts argued that the householder who maintains his sacred fire, procreates children, and performs his ritual duties well also earns religious merit. Nearly 2,000 years ago, these dharma texts elaborated the social doctrine of the four ashramas (stages of life). This concept is an attempt at harmonizing the conflicting tendencies of Hinduism into one system. It held that a member of the three higher classes should first become a chaste student (brahmachari); then become a married householder (grihastha), discharging his debts to his ancestors by begetting sons and to the gods by sacrificing; then retire (as a vanaprastha), with or without his wife, to the forest to devote himself to spiritual contemplation; and finally, but not mandatorily, become a homeless wandering ascetic (sannyasin). The situation of the forest dweller was always a delicate compromise that remained problematic on the mythological level and was often omitted or rejected in practical life.

The householder

Although the status of a householder was often extolled, and some authorities, regarding studentship as a mere preparation, went so far as to brand the other stages as inferior, there were always people who became wandering ascetics immediately after studentship. Theorists were inclined to reconcile the divergent views and practices by allowing the ascetic way of life to those who are, owing to the effects of restrained conduct in former lives, entirely free from worldly desire, even if they had not gone through the traditional prior stages.

THREE MARGAS: PATHS TO SALVATION

Hindus disagree about the way (marga) to final emancipation (moksha). Three paths to salvation (variously valued but nonexclusive) are presented in an extremely influential religious text, the *Bhagavadgītā* ("Song of the Lord"; *c.* 200 BC), according to which it is not acts themselves but the desire for their results that produces karma and thus attachment. These three ways to salvation are (1) the karma-marga ("the path of duties"), the disinterested discharge of ritual and social obligations; (2) the jnana-marga ("the path of knowledge"), the use of meditative concentration preceded by a long and systematic ethical and contemplative training, yoga, to gain a supra-intellectual insight into one's identity with *brahman;* and (3) the bhakti-marga ("the path of devotion"), the devotion to a personal God. These ways are regarded as suited to various types of people.

Although the search for moksha has never been the goal of more than a small minority of Hindus, liberation was a religious ideal that affected all lives. Moksha determined not only the hierarchical values of Indian social institutions and religious doctrines and practices but also the function of Indian philosophy, which is to discuss what one must do to find true fulfillment and what one has to realize, by direct experience, in order to escape from samsara (bondage) and obtain spiritual freedom. While those who have not been reached by formal Indian philosophy have only vague ideas about the doctrines of karma and moksha, in semipopular milieus these doctrines gave rise to much speculation.

For the ordinary Hindu, the main aim of worldly life lies in conforming to social and ritual duties, to the traditional rules of conduct for one's caste, family, and profession.

Dharma: comprehensive system of duties, morals, and religious law

Such requirements constitute an individual's dharma (law and duties), one's own part of the broader stability, law, order, and fundamental equilibrium in the cosmos, nature, and society. *Sanātana* (traditional) dharma—a term used by Hindus to denote their own religion—is a close approximation to "religious practices" in the West. This traditional dharma applies theoretically to all Hindus, but it is superseded by the more particular dharmas that are appropriate to each of the four major varnas, or classes of society: Brahmans (priests), Kṣatriyas (warrior kings), Vaiśyas (the common people), and Śūdras (servants). These four rather abstract categories are further superseded by the more practically applicable dharmas appropriate to each of the thousands of particular castes (*jātis*). Thus, religion for Hindus is mainly a tradition and a heritage, a way of life and a mode of thought. In practice, it is the right application of methods for securing both welfare in this life and a good condition in the hereafter. (W.Do.)

The history of Hinduism

The history of Hinduism began in India about 1500 BC. Although its literature can be traced only to before 1000 BC, evidence of Hinduism's earlier antecedents is derived from archaeology, comparative philology, and comparative religion.

SOURCES OF HINDUISM

Indo-European sources. The earliest literary source for the history of Hinduism is the Rigveda (Ṛgveda), the hymns of which were chiefly composed during the last two or three centuries of the 2nd millennium BC. The religious life reflected in this text is not that of Hinduism but of an earlier sacrificial religious system, generally known as Brahmanism or Vedism, which developed in India among Aryan invaders. This branch of a related group of nomadic and seminomadic tribal peoples originally inhabiting the steppe country of southern Russia and Central Asia brought with them the horse and chariot and the Sanskrit language. Other branches of these peoples penetrated into Europe, bringing with them Indo-European languages that developed into the chief language groups now spoken there.

The religion of the Rigveda

Before they entered the Indian subcontinent (*c.* 1500 BC), the Aryans were in close contact with the ancestors of the Iranians, as evidenced by similarities between Sanskrit and the earliest surviving Iranian languages. Thus, the religion of the Rigveda contains elements from three evolutionary strata: an early element common to most of the Indo-European tribes; a later element held in common with the early Iranians; and an element acquired in the Indian subcontinent itself, after the main Aryan migrations. Hinduism arose from the continued accretion of further elements derived from the original non-Aryan inhabitants, from outside sources, and from the geniuses of individual reformers at all periods.

Hinduism has a few direct survivals from its Indo-European heritage. Some of the rituals of the Hindu wedding ceremony, notably the circumambulation of the sacred fire and the cult of the domestic fire itself, have their roots in the remote Indo-European past. The same is probably true of the custom of cremation and some aspects of the ancestor cult. The Rigveda contains many other Indo-European elements, such as the worship of male sky gods with sacrifices and the existence of the old sky god Dyaus, whose name is cognate with those of the classical Zeus of Greece and Jupiter of Rome ("Father Jove"). The Vedic heaven, the "world of the fathers," resembled the Germanic Valhalla and seems also to be an Indo-European inheritance.

Cremation and sacrifice

Indo-Iranian sources. The Indo-Iranian element in later Hinduism is chiefly found in the initiatory ceremony (*upanayana*) performed by boys of the three upper classes, a rite both in Hinduism and in Zoroastrianism that involves the tying of a sacred cord. The Vedic god Varuṇa, now an unimportant sea god, appears in the Rigveda as sharing many features of the Zoroastrian Ahura Mazdā ("Wise Lord"); the hallucinogenic sacred drink soma corresponds to the sacred *haoma* of Zoroastrianism.

Indigenous sources. Even in the earlier parts of the Rigveda the religion had already acquired numerous specifically Indian features. Some of the chief gods, for example, have no clear Indo-European or Indo-Iranian counterparts. Although some of the new features may have evolved entirely within the Aryan framework, it is generally presumed that many of them stem from the influence of the indigenous inhabitants. The Vedic Aryans may never have been in direct contact with the civilization of the Indus Valley in its prime, but the religion of the valley's culture undoubtedly influenced them.

Non-Indo-European sources. *The Dravidian hypothesis.* Features of Hinduism that cannot be traced to the Rigveda are sometimes ascribed to the influence of the original inhabitants, who are often vaguely and incorrectly referred to as "Dravidians." The ruling classes of the Harappā culture (*c.* 2500–1700 BC), or the Indus civilization, may have spoken a Dravidian language, but as long as their script remains undeciphered this cannot be proved. Moreover, the presence of Dravidian speakers throughout the whole subcontinent at any time in history is not attested. The Mediterranean racial type, to which most modern higher-caste Dravidian speakers belong, is widespread throughout India; but it cannot be proved that all people of this type originally spoke Dravidian languages or that all followed the same culture. Equally or more widely spread in South and Southeast Asia is the Proto-Australoid racial type, the purest members of which in India are the tribal peoples of the centre and the south, many of whom speak languages of the Austric family. Thus, although many aspects of Hinduism are traceable to non-Aryan influence, not all of these aspects are borrowed from "Dravidians." In the 20th century the term Dravidian generally refers to a family of languages and not to an ethnic group.

Dravidian culture

Other sources. The Central Asian nomads who entered India in the two centuries before and after the beginning of the Christian Era might have influenced the growth of devotional Hinduism out of Vedic religion. The classical Western world directly affected Hindu religious art, and several features of Hinduism can be traced to Zoroastrianism. The influence of later Chinese Taoism on Tantric Hinduism (an esoteric system of rituals for spiritual power) has been suggested, though not proved. In more recent centuries, the influence of Islām and Christianity on Hinduism can be seen.

The process of "Sanskritization." The development of Hinduism can be interpreted as a constant interaction between the religion of the upper social groups, represented by the Brahmans (priests and teachers), and the religion of other groups. From the time of the Aryan invasion (*c.* 1500 BC) the indigenous inhabitants of the subcontinent have tended to adapt their religious and social life to Brahmanic norms. This has developed from the desire of lower-class groups to rise on the social ladder by adopting the ways and beliefs of the higher castes. This process, sometimes called "Sanskritization," began in Vedic times when non-Aryan chieftains accepted the ministrations of Brahmans and thus achieved social status for themselves and their subjects. It was probably the principal method by which Hinduism spread through the subcontinent and into Southeast Asia. Sanskritization still continues in the form of the conversion of tribal groups, and it is reflected by the persistent tendency of low-caste Hindus to try to raise their status by adopting high-caste customs, such as wearing the sacred cord and becoming vegetarians.

If Sanskritization has been the main means of spreading Hinduism throughout the subcontinent, its converse process, which has no convenient label, has been one of the means whereby Hinduism has changed and developed over the centuries. The Aryan conquerors lived side by side with the indigenous inhabitants of the subcontinent, and many features of Hinduism, as distinct from Vedic religion, may have been adapted from the religions of the non-Aryan peoples of India. The phallic emblem of the god Śiva arose from a combination of the phallic aspects of the Vedic god Indra and a non-Vedic icon of early popular fertility cults. Many features of Hindu mythology and several of the lesser gods—such as Gaṇeśa, an elephant-headed god, and Hanumān, the monkey god—were

Contributions from non-Aryan religions

incorporated into Hinduism and assimilated into the appropriate Vedic gods by this means. Similarly, the worship of many goddesses who are now regarded as the consorts of the great male Hindu gods, as well as the worship of the one great goddess herself, may have originally incorporated the worship of non-Aryan local goddesses. Unorthodox circles on the fringes of Brahmanic culture (probably in southern India) were one of the important sources of the system of ecstatic devotional religion known as bhakti.

Thus, the history of Hinduism can be interpreted as the imposition of orthodox custom upon wider and wider ranges of people and, complementarily, as the survival of features of non-Aryan religions that gained strength steadily until they were adapted by the Brahmans.

THE PREHISTORIC PERIOD (3RD AND 2ND MILLENNIA BC)

Indigenous prehistoric religion. The prehistoric culture of the Indus Valley arose in the latter centuries of the 3rd millennium BC from the metal-using village cultures of the region. There is considerable evidence of the religious life of the Indus people, but until their writing is deciphered its interpretation is speculative. Enough evidence exists, however, to show that several features of later Hinduism had prehistoric origins.

In most of the village cultures, small terra-cotta figurines of women, found in large quantities, have been interpreted as icons of a fertility deity whose cult was widespread in the Mediterranean area and in western Asia from Neolithic times onward. This hypothesis is strengthened by the fact that the goddess was apparently associated with the bull—a feature also found in the ancient religions farther west.

The Harappā culture

Religion in the Indus Valley civilization. The Harappā culture (often called the Indus Valley civilization), located in modern Pakistan, has produced much evidence of the cult of the goddess and the bull. Figurines of both occur, with the goddess being more common than the bull. The bull, however, appears more frequently on the many steatite seals. A horned deity, possibly with three faces, occurs on a few seals, and on one seal he is surrounded by animals. A few male figurines in hieratic (sacerdotal) poses and one apparently in a dancing posture may represent deities. No building has been discovered at any Harappan site that can be positively identified as a temple, but the Great Bath at Mohenjo-daro was almost certainly used for ritual purposes, as were the ghats (bathing steps on riverbanks) attached to later Hindu temples. The presence of bathrooms in most of the houses and the remarkable system of covered drains indicate a strong concern for cleanliness that may have been related to concepts of ritual purity as well as to ideas of hygiene.

Sacred trees

Many seals show religious and legendary themes that cannot be interpreted with certainty. There is clear evidence, however, of the worship of sacred trees or of the divinities believed to reside in them. The bull is often depicted standing before a sort of altar, and the horned god has been interpreted, perhaps overconfidently, as a prototype of the Hindu god Śiva. Small conical objects appear to be phallic emblems that are also connected with Śiva in later Hinduism, although they may have been pieces used in board games. Other interpretations of the remains of the Harappā culture are more speculative and, if accepted, would indicate that many features of later Hinduism were already in existence 4,000 years ago. The fact that Harappans buried their dead with grave deposits, a practice not followed by the later Hindus, suggests that they had some belief in an afterlife.

Survival of archaic religious practices. Some elements of the religious life of current and past folk religions—notably sacred animals, sacred trees, especially the pipal (*Ficus religiosa*), and the use of small figurines for cult purposes—are found in all parts of India and may have been borrowed from pre-Aryan civilizations. On the other hand, these figures are also commonly encountered outside of India, and therefore they may have originated independently in Hinduism as well.

THE VEDIC PERIOD (2ND MILLENNIUM–7TH CENTURY BC)

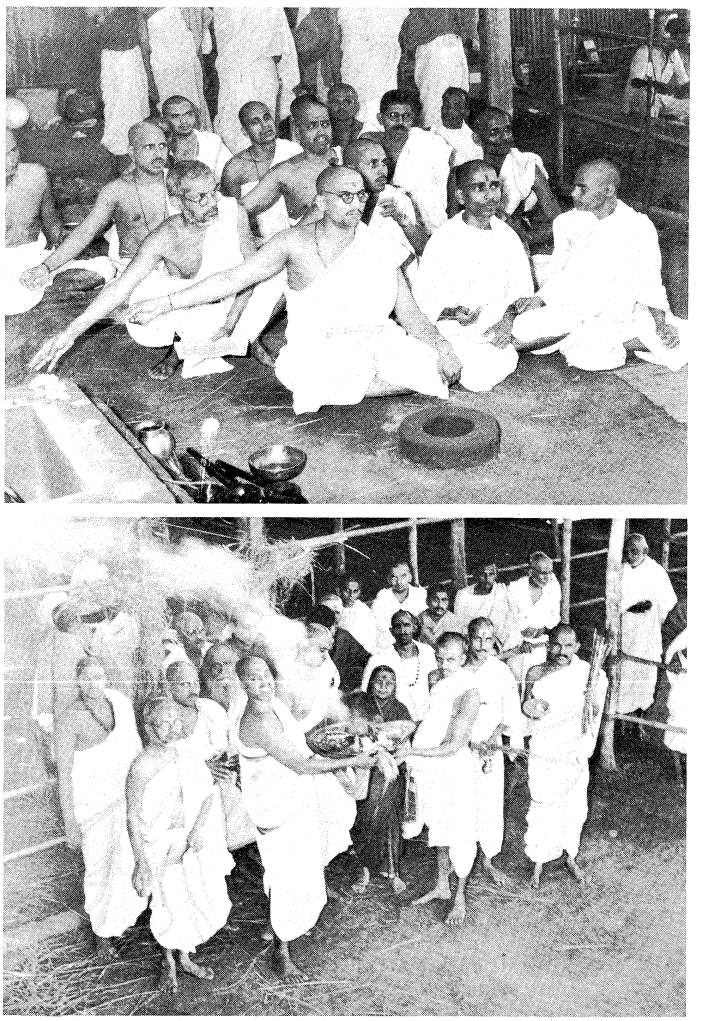
The Aryans of the early Vedic period left few material remains, but they left a very important literary record called the Rigveda. Its 1,028 hymns are distributed throughout 10 books, of which the first and the last are the most recent. A hymn usually consists of three sections: it begins with an exhortation that is followed in the main part by praise of the deity, prayers, and imploration, with frequent references to the deity's mythology, and finishes with a specific request.

The Rigveda ("Wisdom of the Verses") is not a unitary work, and its composition may have taken several centuries. In its form at the time of its final edition it reflects a well-developed religious system. The date commonly given for the final recension of the Rigveda is 1000 BC. During the next two or three centuries the Rigveda was supplemented by three other Vedas and, still later, by Vedic texts called the Brāhmaṇas and the Upanishads (see below *Sacred texts: Vedas*).

CHALLENGES TO BRAHMANISM (7TH–2ND CENTURY BC)

The century from about 550 BC onward was a period of great change in the religious life of India. This century saw the rise of breakaway sects of ascetics who denied the authority of the Vedas and of the Brahmans and who followed founders claiming to have discovered the secret of obtaining release from transmigration. By far the

C.M. Natu



Aspects of a soma sacrifice performed in Pune (Poona), India, on behalf of a traditional Brahman, following the same ritual used in 500 BC, an unusual example of the continuance of the Vedic tradition. (Though Vedic rites are still occasionally performed, photographic documentation of a complete and ritually correct ceremony is rare.) (Top) Group of 16 priests (four for each of the four Vedas) swear to the common interest of the sacrificer before the beginning of the ritual, their hands outstretched toward the fire (*āhavaniya*). (Bottom) Priests bring in fire to light the sacrificial fire in the newly constructed altar, taken from the household fire of the sacrificer, shown with his wife in the middle of the group.

most important of these were Siddhārtha Gautama, called the Buddha, and Vardhamāna, called Mahāvīra ("Great Hero"), the great teacher of Jainism (see also BUDDHISM, THE BUDDHA AND; JAINISM). There were many other heterodox teachers who organized bands of ascetic followers, and each group followed a specific code of conduct. They gained considerable support from ruling families and merchants. The latter were growing in wealth and influence, and many of them were searching for alternative forms of religious activity that would give them a more significant role than did orthodox Brahmanism or that would be less expensive to support.

The scriptures of the new religious movements throw some light on the popular religious life of the period. The god Prajāpati was widely believed to be the highest god and the creator of the universe, with Indra, known chiefly as Śakra ("the Mighty One"), second to him in importance. The Brahmans were very influential, but opposition had developed to their large-scale animal sacrifices—on both philosophical and economic grounds—and their pretensions to superiority by virtue of their birth were questioned. The doctrine of transmigration was by then generally accepted, although a group of outright materialists denied the survival of the soul after death. The ancestor cult, part of the Indo-European heritage, was retained almost universally, at least by the higher castes. Popular religious life largely centred around the worship of local fertility divinities (yaksha), snake-spirits (naga), and other minor spirits in sacred places and groves (*caitya*). Although these sacred places were the main centres of popular religious life, there is no evidence of any buildings or images associated with them, and it appears that neither temples nor large icons existed at the time.

Around 500 BC asceticism became widespread, and increasing numbers of intelligent young men "gave up the world" to search for release from transmigration by achieving a state of psychic security. The orthodox Brahmanical teachers reacted to these tendencies by devising the doctrine of the four ashramas (*āśramas*, "abodes"), which divided the life of the twice-born after initiation into four stages: the brahmachari (celibate religious student); the grihastha (married householder); the vanaprastha (forest dweller); and the sannyasin (wandering ascetic). This attempt to keep asceticism in check and confine it to men of late middle age was never followed universally, but thereafter Hindu social theory centred on the concept of varnashramadharma, or the duties of the four classes (varna) and the four stages of life (ashramas), which formed the ideal that Hindus were encouraged to follow.

The doctrine of the four ashramas

The 3rd century BC was the period of the Mauryan empire, the first great empire of India. Its early rulers were heterodox, and Aśoka (reigned *c.* 265–238 BC), the third and most famous of the Mauryan rulers, was a professed Buddhist. Although there is no doubt that Aśoka's patronage of Buddhism did much to spread that religion, his inscriptions recognize the Brahmans as worthy of respect. Sentiments in favour of nonviolence (ahimsa) and vegetarianism, much encouraged by the heterodox sects, spread during the Mauryan period and were greatly encouraged by Aśoka. A Brahmanic revival appears to have occurred with the fall of the Mauryas. The orthodox religion itself was undergoing change at this time, however, because of the development of theistic tendencies that centred around the gods Vishnu and Śiva.

Inscriptions, iconographic evidence, and literary references point to the emergence of devotional theism in the 2nd century BC. Several brief votive inscriptions refer to the god Vāsudeva, who by this time was widely worshiped in western India. At the end of the 2nd century, Heliodorus, a Greek ambassador from King Antialcidas of Taxila (in Pakistan), erected a large column in honour of Vāsudeva at Besnagar in Madhya Pradesh and recorded that he was a Bhāgavata, a term specifically used for the devotees of Vishnu. The identification of Vāsudeva with the old Vedic god Vishnu and, later, with Vishnu's incarnation, Krishna (Kṛṣṇa), was quickly accepted.

The beginning of devotional theism

Near the end of the Mauryan period the first surviving stone images of Hinduism appear. Several large, simply carved figures survive, not representing any of the great gods but rather yakshas, or local chthonic divinities connected with water, fertility, and magic. The original locations of these images are uncertain, but they were probably erected in the open air in sacred enclosures. Temples are not clearly attested in this period by either archaeology or literature. A few fragmentary images thought to be those of Vāsudeva and Śiva, the latter in anthropomorphic form and in the form of a lingam, or phallic emblem, are found on coins of the 2nd and 1st centuries BC.



Brahman priest reading the unbound folio manuscript of a sacred text, a *śrauta-sūtra*, at a Vedic *yajña* (sacrifice). The animal skin is the sacrificer's seat, the wooden cups are for drinking the soma.
C.M. Natu

EARLY HINDUISM (2ND CENTURY BC–4TH CENTURY AD)

The centuries immediately preceding and following the dawn of the Common Era saw the recension of the two great Sanskrit epics, the *Rāmāyaṇa* and the *Mahābhārata* (the latter incorporating into it the *Bhagavadgītā*). Although it was the worship of Vishnu, incarnate as Krishna in the *Mahābhārata* and as Rāma in the *Rāmāyaṇa*, that developed significantly during this period (see below *Sacred texts: Epics and Purāṇas*), the god Śiva is active in the *Mahābhārata*, and the cult of Śiva developed alongside the cult of Vishnu.

The rise of the major sects: Vaishnavism, Śaivism, and Śāktism. The Vedic god Rudra gained in importance from the end of the Rigvedic period. In the Svetāśvatara Upanishad, Rudra is for the first time called Śiva and is described as the creator, preserver, and destroyer of the universe. His followers are called on to worship him with devotion (bhakti). The tendency for the laity to form themselves into religious guilds, or societies—evident in the case of the yaksha cults, Buddhism, and Jainism—promoted the growth of devotional Vaishnavism and Śaivism. These local associations of worshipers appear to have been a principal factor in the spread of the new cults. Theistic ascetics are less in evidence at this time; but a community of Śaivite monks, the Pāśupatas, was also in existence by the 2nd or 3rd century AD.

The period between the fall of the Mauryan empire (*c.* 185 BC) and the rise of the Gupta (*c.* AD 320) was one of great change, with most of the area of Pakistan and parts of western India being conquered by a succession of invaders. India was opened to influence from the West as never before, not only by its invaders but by way of the sea through the flourishing trade with the Roman Empire. The effects of the new contacts were most obvious in art and architecture. The oldest freestanding stone temple in the subcontinent has been excavated at Taxila, near Rāwalpindi, Pak. During the 1st century BC the Gandhāra school of sculpture arose in the same region and made use of Hellenistic and Roman prototypes, mainly in the service of Buddhism. At that time Hindu temples proba-

Western influences

bly were made of wood because no remains of them are extant; however, literary evidence shows that they must have existed.

By the time of the early Gupta empire the new theism had been harmonized with the old Vedic religion, and two of the main branches of Hinduism were fully recognized. The Vaishnavas had the support of the Gupta emperors, who took the title *paramabhāgavata* ("supreme devotee of Vishnu"). Vishnu temples were numerous and the doctrine of Vishnu's avatars (incarnations) was widely accepted. Of the 10 incarnations of later Vaishnavism, however, only two seem to have been much worshiped in the Gupta period. These were Krishna, the hero of the *Mahābhārata,* who also begins to appear in his pastoral aspect as the cowherd and flute player, and the divine boar (Varāha), of whom several impressive images survive from the Gupta period.

The Śaivites were also a growing force in the religious life of India. The sect of Pāśupata ascetics, founded by Lakulīśa (or Nahulīśa), who lived in the 2nd century AD, is attested by inscriptions from the 5th century and is among the earliest of the sectarian religious orders of Hinduism. Representations of the son of Śiva, Skanda (also called Kārttikeya, the war god), appeared on Kushān coins as early as AD 100. Śiva's other son, the elephant-headed Gaṇeśa, patron deity of commercial and literary enterprises, did not appear until the 5th century. Very important in this period was Sūrya, the sun god, who had temples built in his honour, although in modern times he is little regarded by most Hindus. The solar cult had Vedic roots but later may have expanded under Iranian influence.

Rising importance of goddesses

Several goddesses began to gain importance in this period. Although goddesses had always been worshiped in local and popular cults, they play comparatively minor roles in Vedic religion. Lakṣmī or Śrī, goddess of fortune and consort of Vishnu, was worshiped before the beginning of the Christian era, and several lesser goddesses are attested from the Gupta period. But the cult of Durgā, the consort of Śiva, was only beginning to gain importance in the 4th century, and the large-scale development of Śāktism (devotion to the active, creative principle personified as the Mother Goddess) did not take place until medieval times.

The development of temples. The Gupta period (4th–6th century) saw the rapid development of temple architecture. Earlier temples were made of wood, but free-standing stone and brick temples soon appeared in many parts of India. By the 7th century, stone temples, some of considerable dimensions, were found in the Aryanized parts of the country. Originally, the design of the Hindu temple may have borrowed from the Buddhist precedent, for in some of the oldest temples the image was placed in the centre of the shrine, which was surrounded by an ambulatory path resembling the path around the Buddhist stupa (a religious building containing a relic). Nearly all surviving Gupta temples are comparatively small; they consist of a small cella (central chamber), constructed of thick and solid masonry, with a veranda either at the entrance or on all sides of the building. The earliest Gupta temples, such as the Buddhist temples at Sānchi, have flat roofs; however, the *sikhara* (spire), typical of the north Indian temple, was developed in this period and with time steadily was made taller. The massive and tall tower of the Buddhist temple of Buddh Gayā, which was in existence in the 7th century, represents the culmination of Gupta temple architecture.

The Buddhists and Jains had made use of artificial caves for religious purposes, and these were adapted by the Hindus. Hindu cave-temples, however, are comparatively rare, and none has been discovered from earlier than the Gupta period. In the Pallava site of Mahābalipuram, south of Madras, a number of small temples were carved in the 7th century from outcroppings of rock and represent some of the oldest religious buildings in the Tamil country.

The spread of Hinduism in Southeast Asia and the Pacific. Hinduism and Buddhism had an immense impact on the civilizations of Southeast Asia and contributed greatly to the development of a written tradition in that area. Around the beginning of the Christian era, Indian merchants in comparatively large numbers settled there, bringing Brahmans and Buddhist monks with them. These religious men were patronized by local chiefs, who converted to the new religion. The earliest material evidence of Hinduism in Southeast Asia comes from Borneo, where late 4th-century Sanskrit inscriptions testify to the performance of Vedic sacrifices by Brahmans at the behest of local chiefs. Chinese chronicles attest an Indianized kingdom in Vietnam two centuries earlier. The dominant form of Hinduism exported to Southeast Asia was Śaivism, though some Vaishnavism was also known there. Later, from the 9th century onward, Tantrism, both Hindu and Buddhist, spread throughout the region.

The civilizations of Southeast Asia developed forms of Hinduism and Buddhism that had distinctive local features and were attuned to the local cultures, but the framework of their religious life was essentially Indian. Stories from the *Rāmāyaṇa* and the *Mahābhārata* became widely known in Southeast Asia and are still popular there in local versions. The people of Bali (in Indonesia) still follow a form of Hinduism adapted to their own genius. Versions of the *Manu-smṛti* ("Laws of Manu") were taken to Southeast Asia and were translated and adapted to indigenous cultures until they lost most of their original content.

Claims of early Hindu contacts farther east are more doubtful. There is little evidence of the influence of Hinduism on China and Japan, except through Buddhism.

Indian religious influence in the Mediterranean world. Nearly as dubious as the question of Hindu influence on the religious life of the Far East is its influence on that of the ancient Mediterranean world. The Greek philosopher Pythagoras may have obtained his doctrine of metempsychosis (transmigration, or passage of the soul from one body to another) from India, mediated by Achaemenian (6th–4th century BC) Persia, but similar ideas were known in Egypt and were certainly present in Greece before the time of Pythagoras. The Pythagorean doctrine of a cyclic universe may also be derived from India, but the Indian theory of cosmic cycles is not attested in the 6th century BC. Nevertheless, it is known that Hindu ascetics occasionally visited Greece. The most striking similarity of Greek and Indian thought is the resemblance between the system of mystical gnosis (esoteric knowledge) described in the *Enneads* of the Neoplatonic philosopher Plotinus (3rd century AD) and that of the *Yoga-sūtras* attributed to Patañjali, an Indian religious teacher sometimes dated in the 2nd century AD. The Patañjali text is the older, and influence must be suspected, though the problem of mediation remains difficult because Plotinus gives no direct evidence of having known anything about Indian mysticism. Several Greek (*e.g.,* Clement of Alexandria) and Latin writers show considerable knowledge of the externals of Indian religions, but none gives any intimation of understanding their more recondite aspects.

Certain Vaishnava legends, especially those referring to the infant Krishna, bear some resemblance to those of Christianity, and claims have been made by both Hinduism and Christianity that the one influenced the other. There is, however, no definitive evidence for the priority of either one.

THE RISE OF DEVOTIONAL HINDUISM (4TH–11TH CENTURY)

The medieval period saw the growth of new devotional religious movements centred on hymnodists who taught in the popular languages of the time. The new movements probably began with the appearance of hymns in Tamil associated with two groups of poets, the Nāyaṇārs, worshipers of Śiva, and the Āḻvārs, devoted to Vishnu. The oldest of these date from the early 7th century, although passages of devotional character can be found in earlier strata of Tamil literature.

Tamil influences

The term bhakti, in the sense of devotion to a personal god, appears in the *Bhagavadgītā* and the Śvetāśvatara Upanishad. In these early sources it represents a devotion still somewhat restrained and unemotional. The new form of bhakti, associated with singing in the languages of the common people, was highly charged with emotion, and the relation of worshiper and divinity was often de-

scribed by the analogy of that of lover and beloved. This devotional poetry is characterized by a mystical fervour not found in the Upanishads and the *Bhagavadgītā,* in both of which, even when the object of meditation is conceived as a personal God, there is little expression of passion. The Tamil "saints," however, felt an intense love (Tamil: *anbu*) of a personal kind toward their god. They experienced overwhelming joy in his presence and deep sorrow when he did not reveal himself. Some of them felt a profound sense of guilt or inadequacy in the face of the divine. But the dominant emotion in these poems is one of joy, often expressing itself in song and dance. The poems have a strong ethical content and encourage the virtues of love, humility, and brotherhood. The ideas of these poets, spreading northward, probably were the origin of the growth of bhakti in northern India.

Antagonism between devotional cults and Buddhism

The devotional cults further weakened Buddhism, which had long been on the decline. From time to time Hindus, especially Śaivites, took aggressive action against Buddhism. At least two Śaivite kings—the Hephthalite invader Mihirakula (early 6th century) and the Bengal king Śaśāṅka (early 7th century)—are reported to have been active persecutors, destroying monasteries and killing monks. The philosophers Kumārila and Śaṅkara were also strongly opposed to Buddhism. In their journeys throughout India, their biographies claim, they vehemently debated with Buddhists and tried to persuade kings and other influential people to withdraw their support from Buddhist monasteries. Only in Bihār and Bengal, because of the patronage of the Pāla dynasty and some lesser kings and chiefs, did Buddhist monasteries continue to flourish. Buddhism in eastern India, however, was well on the way to being reabsorbed into Hinduism when the Muslims invaded the Ganges (Gaṅgā) Valley in the 12th century. The great Buddhist shrine of Buddh Gayā, the site of the Buddha's enlightenment, became a Hindu temple and remained as such until recent times.

At the end of its existence in India, Buddhism developed in a way that had some effect on Hinduism. Among the Buddhist Tantrists appeared a new school of preachers, often known as siddhas (those who have achieved), who sang their verses in the contemporary languages, early Maithili and Bengali. They taught that giving up the world was not necessary for release from transmigration and that by living a life of simplicity in one's own home one could achieve the highest state. This system, known as Sahajayāna ("the Vehicle of the Natural," or "the Easy Vehicle"), influenced both Bengali devotional Vaishnavism, which produced sects called Sahājiyā with similar doctrines, and the Nātha yogis (mentioned below), whose teachings influenced Kabīr and other later bhakti teachers.

HINDUISM UNDER ISLĀM (11TH–19TH CENTURY)

The challenge of Islām and popular religion. The phase of Indian history marked by the domination of the Muslims in most of northern India saw great changes in Indian religion. The advent of Islām in the Ganges Basin at the end of the 12th century resulted in the withdrawal of royal patronage from Hinduism in much of the area. The attitude of the Muslim rulers toward Hinduism varied. Some, like Fīrūz Tughluq (ruled 1351–88) and Aurangzeb (ruled 1658–1707), were strongly anti-Hindu and enforced payment of jizya, a poll tax on unbelievers. Others, like the Bengali sultan Ḥusayn Shāh 'Alā' ad-Dīn (reigned 1493–1519) and the great Akbar (reigned 1556–1605), were well-disposed toward their Hindu subjects. Many temples, however, were destroyed by the more fanatical rulers. Conversion to Islām was more common in areas where Buddhism had once been strongest—modern Pakistan, Bangladesh, and Kashmir.

Hinduism on the eve of the Muslim occupation

On the eve of the Muslim occupation, Hinduism was by no means sterile in northern India, but its vitality was centred in the southern, Dravidian-speaking areas. Throughout the centuries, the system of class and caste had become more rigid; in each region there was a complex hierarchy of castes strictly forbidden to intermarry and interdine, controlled and regulated by secular powers who acted on the advice of the court Brahmans. The large-scale Vedic sacrifices had practically vanished, but simple domestic Vedic sacrifices continued, and new forms of animal, and sometimes vegetarian, sacrifice had appeared, especially connected with the cult of the Mother Goddess.

By that time, the main divinities of later Hinduism were worshiped. Rāma, the hero of the epic poem, had become the eighth avatar of Vishnu, and his cult was growing, although it was not yet as prominent as it later became. Similarly, Rāma's monkey helper, Hanumān, now one of the most popular divinities of India and the most ready helper in time of need, was rising in importance. Krishna was worshiped with his adulterous consort, Rādhā. Strange syncretic gods had appeared, such as Harihara, a combination of Vishnu and Śiva, and Ardhanārīśvara, a synthesis of Śiva and his shakti Pārvatī or Durgā.

Hindu temples

Temple complexes. From the Gupta period onward Hindu temples tended to become larger and more prominent, and their architecture developed in distinctive regional styles. In northern India the best remaining Hindu temples are found in the Orissa region and in the town of Khajurāho in northern Madhya Pradesh. The best example of Orissan temple architecture is the Liṅgarāja temple of Bhubaneswar, built about 1000. The largest temple of the region, however, is the famous Black Pagoda, the Sun Temple (Sūrya Deuḷa) of Konārak, built in the mid-13th century. Its tower has long since collapsed, and only the assembly hall remains. The most important Khajurāho temples were built during the 11th century. Individual architectural styles also arose in Gujarāt and Rājasthān, but their surviving products are less impressive than those of Orissa and Khajurāho. By the end of the 1st millennium AD the South Indian style had reached its apogee in the great Rājarājeśvara temple of Thanjāvūr (Tanjore).

In the temple the god was worshiped by the rites of puja (reverencing a sacred being or object) as though the worshipers were serving a great king. In the important temples a large staff of trained officiants waited on the god. He was awakened in the morning along with his goddess, washed, clothed and fed, placed in his shrine to give audience to his subjects, praised and entertained throughout the day, ceremoniously fed, undressed, and put to bed at night. Worshipers sang, burned lamps, waved lights before the divine image, and performed other acts of homage. The god's dancing girls (devadasis) performed before him at regular intervals, watched by the officiants and lay worshipers, who were his courtiers. These women, either the daughters of devadasis or girls dedicated in childhood, may have also served as prostitutes. The association of dedicated prostitutes with certain Hindu shrines can be traced back to before the Christian era. It became more widespread in post-Gupta times, especially in South India, and aroused the reprobation of 19th-century Europeans. Through the efforts of Hindu reformers the office of the devadasis was discontinued. The role of devadasis is best understood in the context of the analogy between the temple and the royal court, for the Hindu king also had his

By courtesy of a private collector, Cambridge, Mass.



"Four Priests Doing Puja Before the Image of Śrī Nāthajī," Rajasthani miniature painting from Nāthdwāra, *c.* 1830. In the Fogg Art Museum, Cambridge, Mass.

dancing girls, who bestowed their favours on his courtiers.

Parallels between the temple and the royal palace also were in evidence in the *rathayātra*s (shrine processions). As on festival days, when the king issued from his palace and paraded around his city, escorted by courtiers, troops, and musicians, so also the god paraded around his city in a splendid procession, together with the lesser gods of the minor shrines. The god rode on a tremendous and ornate moving shrine (*ratha*), which was often pulled by large bands of devotees. *Rathayātra*s still take place in many cities of India. The best-known is the annual procession of Jagannātha ("Juggernaut"), a form of Vishnu, at Puri, Orissa.

The wealth of the great temples

The great temples were (and still are) wealthy institutions. They were supported by the transfer of the taxes levied by kings on specific areas of the nearby countryside, by donations of the pious, and by the fees of worshipers. Their immense wealth was one of the factors that encouraged the Ghaznavid and Ghūrid Turks to invade India after the 11th century. They were controlled by self-perpetuating committees—whose membership was usually a hereditary privilege—and by a large staff of priests and temple servants under a high priest who wielded tremendous power and influence. The great walled temple complexes of South India were (and still are) small cities, containing the central and numerous lesser shrines, bathing tanks, administrative offices, homes of the temple employees, workshops, bazaars, and public buildings of many kinds. Directly and indirectly they played an important part in the economy, as they were among the largest employers and greatest landowners in their areas. They also performed valuable social functions because they served as schools, dispensaries, poorhouses, banks, and concert halls.

The Muslim occupation brought India into close contact with a different, more aggressive, religion. In such circumstances, the absence of a central religious authority in Hinduism was a source of strength. The *purohita*s, or family priests who performed the domestic rituals and personal sacraments for the lay people, continued to function, as did the thousands of ascetics. In Muslim-occupied territory the temples suffered the most. In the sacred cities of Vārānasi (Benares) and Mathurā, no large temple remains from any period before the 17th century. The same is true of most of the main religious centres of northern India, but not of the regions where the Muslim hold was less firm, such as Orissa, Rājasthān, and South India.

Sectarian movements. Before the time the Muslims invaded the subcontinent, the new forms of South Indian bhakti were spreading beyond the bounds of the Dravidian south. Certain Vaishnava theologians of the Pāñcarātra and Bhāgavata schools, including Rāmānuja, a Tamil Brahman who was for a time chief priest of the Vaishnava temple of Srīrangam, near Tiruchchirāppalli (Trichinopoly), taught in the 11th century. They gave the growing Vaishnava bhakti cults a philosophical framework that also influenced some Śaivite schools.

Two other Vaishnava teachers deserve mention. Nimbārka, a Telugu Brahman of the 12th or 13th century, spread the cult of the divine cowherd and his favourite *gopī* (cowherdess, especially associated with the legends of Krishna's youth), Rādhā. His sect survives near Mathurā but has made little impact elsewhere. More important was Vallabha (Vallabhācārya; 1479–1531), who took the Vaishnava doctrine of grace and emphasized its erotic imagery. His sect is noteworthy because it stresses absolute obedience to the guru (teacher). Early in its existence it was organized with a hierarchy of senior monks (*gosvāmī*), many of whom became very rich. The Vallabhācārya sect was once very influential in the western half of North India, but it declined in the 19th century, in part because of a number of lawsuits against the chief guru, the descendant of Vallabha.

Vaishnava and Śaiva teachers

The Śaiva sects also developed from the 10th century onward. In South India there emerged the school of Śaiva-siddhānta, still one of the most significant religious forces in that region, and one that, unlike the school of Śaṅkara, does not admit the full identity of the soul and God. A completely monistic school of Śaivism appeared in Kashmir in the early 9th century. Its doctrines differ from those of Śaṅkara chiefly because it attributes personality to the absolute spirit, who is the god Śiva and not the impersonal *brahman*.

The unorthodoxy of the Liṅgāyats

An important and interesting sect, founded in the 12th century in the Kannada-speaking area of the Deccan, was that of the Liṅgāyats, or Vīraśaivas ("Heroes of the Śaiva Religion"). Its traditional founder, Basava, taught doctrines and practices of surprising unorthodoxy: he opposed all forms of image worship and accepted only the lingam of Śiva as a sacred symbol. Vīraśaivism rejected the Vedas, the Brahman priesthood, and all caste distinction. Several Liṅgāyat practices, now largely abandoned, such as the remarriage of widows and the burial of the dead, are deliberately antinomian.

An important development of Śaivism in North India was brought about by Gorakhnāth (Gorakṣanātha), who in the 13th century became leader of a sect of Śaivite ascetics known as Nātha ("Lord") from the title of their chief teachers. The Gorakhnāthīs were particularly important as propagators of the practices of hatha-yoga, a form of yoga that requires complex and difficult physical exercises and that has become popular in the West. These yogis, who are still numerous, influenced the teaching of several of the bhakti poets.

Bhakti movements. The poets and "saints" of medieval bhakti appeared throughout India. Although all have their individual genius, the bhakti lyricists share a number of common features whatever their language. The Sanskrit education needed for authors of Sanskrit texts limited them largely to the Brahman class and thus put a definite stamp on them. Because bhakti poets could use any language, they might come from any class. They brought to their poetry a familiarity with folk religion unknown or ignored in the Sanskrit texts. The use of the spoken language, even though it was formalized, made possible the immediate expression of an unmediated vision that needed no further context; thus the lyrics are short, intensely personal, and precise. These works illustrate the localistic and reformist tendency evidenced throughout India in the vernacular literatures, especially in Tamil, Bengali, and Hindi. (See below *Vernacular literatures*.)

The origin of the new forms of Hinduism has been attributed to the influence of Islām, but the proposition that the rise of popular emotional bhakti was a response to Islām is impossible, for the practice of singing ecstatic hymns in the current local language was well-known in South India even before Muḥammad. All the features of this form of bhakti are found in the *Bhāgavata-Purāṇa* and in the commentaries of Rāmānuja. The earliest bhakti literature in a living Indo-Aryan language is from Mahārāshtra and was composed before Muslims occupied the area. Thus, passionate bhakti existed long before the Muslim conquest. However, the presence of rulers of alien faith and the withdrawal of royal patronage from the temples and Brahmanic colleges may have encouraged the spread of new, more popular forms of Hinduism. The psychological effect of the Muslim conquest may also have predisposed the people to accept the powerful teachings of the poets, but Islām was only a contributory factor in the spread of the new movements.

Much has been said about the synthesis of Hinduism and Islām in the period of Muslim dominance, but, as far as the Hindus were concerned, this was generally a matter of superficial observances. Thus, purdah (*pardā*), the strict seclusion of women, became commonplace among the Hindu upper classes of northern India, numerous Muslim social customs were adopted, and Persian and Arabic words entered the vocabularies of Indian languages. The fundamental theology of Hinduism, however, was unaffected by Islām, even in the teachings of such men as Basava and Kabīr, who may have been somewhat influenced by Muslim observances and social customs.

The Hindu influence on Islām

What synthesis did take place came from the Muslims, most of whom were Indian by blood. In Hindi, Bengali, Gujarati, Punjabi, and Marathi there is much poetic literature, written by Muslims and commencing with the Islāmic invocation of Allāh, which nevertheless betrays strong Hindu influence. Thus, there are texts that proclaim Krishna as being in the line of the prophets of

Islām and as the teacher of the unity of God. Much mystical poetry, though written by authors with Muslim names, uses Hindu imagery and Hindu terminology. This literature originated in the accommodating character of early Indian Ṣūfism, which, well before Kabīr, proclaimed that Muslim, Christian, Jew, Zoroastrian, and Hindu were all striving toward the same goal and that the outward observances that kept them apart were false. Some of the Indian Ṣūfīs were greatly influenced by Hindu customs. For example, a school of Kashmir Ṣūfīs, whose members call themselves rishis, after the legendary Hindu sages (*ṛṣi*), respect and repeat the verses of Lāl Ded and are strict vegetarians.

Syncretic tendencies were encouraged by tolerant Muslim rulers, and these tendencies reached their zenith in the reign of Akbar (1556–1605), who took a great interest in the religion of his Hindu subjects, favoured vegetarianism, and tried to establish a single, all-embracing religion for his empire. Although the efforts of Akbar failed, they influenced India for more than 50 years after his death. The orthodox Muslim theologians had long been complaining about the growth of heresy, however, and the emperor Aurangzeb (reigned 1658–1707) did all in his power to discourage it. Popular Muslim preachers throughout the 18th and 19th centuries worked to restore orthodoxy. Thus, syncretic tendencies virtually came to an end before the imposition of British power in the mid-18th century. Furthermore, British rule emphasized the distinctions between Hindu and Muslim and did not encourage efforts to harmonize the two religions.

THE MODERN PERIOD (19TH–20TH CENTURY)

Western Christian missionary activities

From their small coastal settlements in southern India, the Portuguese promoted Roman Catholic missionary activity and made converts, most of whom were of low caste; the majority of caste Hindus were unaffected. Small Protestant missions operated from the Danish factories of Tranquebar in Tamil Nādu and Serampore in Bengal, but they were even less influential. The British East India Company, conscious of the disadvantages of unnecessarily antagonizing its Indian subjects, excluded all Christian missionary activity from its territories. Indeed, the company continued the patronage accorded by indigenous rulers to many Hindu temples and forbade its Indian troops to embrace Christianity. The growing evangelical conscience in England brought this policy to an end with the renewal of the company's charter in 1813. The company's policy then became one of strict impartiality in matters of religion, and missionaries were allowed to work throughout its territory. Thus, Christian ideas began to spread.

Hindu reform movements. *Brahmo Samaj.* The pioneer of reform was Rammohan Ray. His intense belief in strict monotheism and in the evils of image worship began early and probably was derived from Islām, because at first he had no knowledge of Christianity. He later learned English and in 1814 settled in Calcutta, where he was prominent in the movement for encouraging education of a Western type. His final achievement was the foundation of the Brahmo Samaj ("Society of God") in 1828.

Rammohan outwardly remained a Hindu, wearing the sacred cord and keeping most of the customs of the orthodox Brahman; but his theology was surprisingly un-Indian. He was chiefly inspired by 18th-century Deism (rational belief in a transcendent creator god) and Unitarianism (belief in God's essential oneness), but some of his writing suggests that he was aware of the religious ideas of the Freemasons (a secret fraternity that espouses some deistic concepts). Several of his friends were members of a Masonic lodge in Calcutta. His ideas of the afterlife are obscure, and it is possible that he did not believe in the doctrine of transmigration. Rammohan was one of the first higher-class Hindus to visit Europe, where he was much admired by the intelligentsia of Britain and France.

Devendranath Tagore and Keshab Chunder Sen

After Rammohan Ray's death, Devendranath Tagore became leader of the Brahmo Samaj, and under his guidance a more mystical note was sounded by the society. The third great leader of the Brahmo Samaj, Keshab Chunder Sen, was a vigorous reformer who completely abolished caste in the *samāj* and admitted women. As his theology became more syncretistic and eclectic, a schism developed, and the more conservative faction remained under the leadership of Tagore. Keshab's faction, the Brahmo Samaj of India, adopted as its scripture a selection of theistic texts gathered from all the main religions; at the same time, it became more Hindu in its worship, employing the *saṃkīrtana* (hymn-singing session) and *nagara-kīrtana* (street procession) of the Caitanya sect. In 1881 Keshab founded the Church of the New Dispensation (Naba Bidhan) for the purpose of establishing the truth of all the great religions in an institution that he believed would replace them all. When he died in 1884, the Brahmo Samaj began to decline, but it produced the greatest poet of modern India, Rabindranath Tagore (1861–1941), son of the second of its great leaders, Devendranath Tagore.

Arya Samaj. A reformer of different character was Dayanand Sarasvati, who was trained as a yogi but steadily lost faith in yoga and many other aspects of Hinduism. After traveling widely as an itinerant preacher, he founded the Arya Samaj in 1875, and it rapidly gained ground in the west of India. Dayanand rejected image worship, sacrifice, and polytheism and claimed to base his doctrines on the four Vedas as the eternal word of God. Later Hindu scriptures were judged critically, and many of them were believed to be completely evil. The Arya Samaj did much to encourage Hindu nationalism, but it did not disparage the knowledge of the West, and it established many schools and colleges. Among its members was the revolutionary Lala Lajpat Rai.

New religious movements. *Ramakrishna Mission.* The most important developments in Hinduism, however, did not arise primarily from the new *samāj*es.

Influence of the mystic Ramakrishna

The mystic Ramakrishna, who was a devotee at a temple of Kālī called Dakṣiṇeśvar to the north of Calcutta, attracted a band of educated lay followers who spread his doctrines. As a result of his studies and visions, he came to the conclusion that "all religions are true" but that the religion of a person's own time and place was for him the best expression of the truth. Even idolatry met the needs of simple people and was not to be disparaged. Ramakrishna thus gave educated Hindus a basis on which they could justify the less rational aspects of their religion to a consciousness increasingly influenced by Western values.

Among the followers of Ramakrishna was Narendranath Datta, who became an ascetic after his master's death and assumed the religious name Vivekananda. In 1893 he attended the World's Parliament of Religions in Chicago, where his powerful personality and stirring oratory deeply impressed the gathering. After lecturing in the United States and England, he returned to India in 1897 with a small band of Western disciples. There he founded the Ramakrishna Mission, the most important modern organization of reformed Hinduism. Vivekananda, more than any earlier Hindu reformer, encouraged social service and the uplift of the downtrodden. Influenced by progressive Western political ideas, he set himself firmly against all forms of caste distinction and fostered a spirit of self-reliance in his followers. The Ramakrishna Mission has done much to spread a knowledge of its version of Hinduism outside India and now has branches in many parts of the world.

Theosophical Society. Another neo-Hindu movement, which at one time exerted considerable influence, is the Theosophical Society. Founded in New York City in 1875 by Helena Blavatsky of Russia, it had as its original inspiration Kabbala (Jewish esoteric mysticism), Gnosticism (esoteric salvatory knowledge), and other forms of Western occultism. When Blavatsky went to India in 1879, her doctrines quickly took on an Indian character, and from her headquarters at Adyar she and her followers established branches in many cities of India.

The society survived serious accusations of charlatanry leveled against its founder and certain other leaders, and it reached the peak of its influence under its next important leader, Annie Besant, a reform-minded Englishwoman. Under her guidance, many Theosophical lodges were founded in Europe and the United States, and these helped to acquaint the West with the principles of Hinduism, if in a rather idiosyncratic form.

Aurobindo Ashram. Another modern teacher whose doctrines have had some influence outside India was Śrī Aurobindo, who began his career as a revolutionary. He later withdrew from politics and settled in Pondicherry, then a French possession. There he established an ashram, or *āśrama* (a retreat), and achieved a high reputation as a sage. His followers saw him as the first incarnate manifestation of the superbeings whose evolution he prophesied, and apparently he did not discourage this belief. After his death, the leadership of the Aurobindo Ashram was assumed by Mira Richard, a Frenchwoman who had been one of his disciples.

Other reform movements. Numerous other teachers have affected the religious life of modern India. Among them was the great Bengali poet Rabindranath Tagore, who was influenced by many currents of earlier religious thought, both Indian and other. Tagore was particularly popular in Europe and America around the time of World War I, and he did much to disseminate Hindu religious thought in the West.

Less important outside India, but much respected in India itself, especially in the Dravidian south, was Ramana Maharshi, a Tamil mystic who maintained almost complete silence. His powerful personality attracted a large band of devotees.

Swami Śivānanda, who had been a physician, established an ashram and an organization called the Divine Life Society near the sacred site of Rishikesh in the Himalayas. This organization has numerous branches in India and some elsewhere. His movement teaches more or less orthodox Vedānta, combined with both yoga and bhakti, but rejects caste and stresses social service.

Jiddu Krishnamurti represents the most attenuated form of Hinduism. He rejects all religious organization and makes no claim to special revelations or exceptional spiritual development; instead, he teaches self-realization through introspection and the abandonment of personal ambition.

Religion and the growth of nationalism

The struggle for independence. The Hindu revival and reform movements of the 19th and early 20th centuries were closely linked with the growth of Indian nationalism and the struggle for independence. The Arya Samaj strongly encouraged nationalism, and even though Swami Vivekananda and the Ramakrishna Mission were always uncompromisingly nonpolitical, their effect in promoting the movement for self-government is quite evident.

Religion and politics were joined in the career of Bal Gangadhar Tilak, an orthodox Mahārāshtrian Brahman who believed that the people of India could be aroused only by appeals couched in religious terms. Tilak used the annual festival of the god Gaṇeśa (Gaṇapati) for nationalist propaganda. His interpretation of the *Bhagavadgītā* as a call to action was also a reflection of his nationalism, and through his mediation the *Bhagavadgītā* became a stimulus to later leaders, including Mahatma Gandhi.

Hindu religious concepts also were enlisted in the nationalist cause in Bengal. In his historical novel *Ānandamaṭh,* the Bengali novelist Bankim Chandra Chatterjee described a band of martial ascetics at the time of the decline of the Mughal empire, who were pledged to free India from Muslim domination. These had as their anthem a stirring devotional song written in simple Sanskrit—"Bande Mātaram" ("I revere the Mother"). The Mother referred to is both the stern demon-destroying goddess Kālī and a personification of India. This song was soon adopted by the more extreme nationalists. Vivekananda emphasized the need to turn the emotion of bhakti toward the suffering poor of India. During his short career as a young revolutionary leader, Śrī Aurobindo made much use of "Bande Mātaram," and he called on his countrymen to strive for the freedom of India in a spirit of devotion. The bhakti of the medieval poets was thus enlisted in the cause of modern independence.

Mahatma Gandhi. Much influenced by the traditional bhakti of his native Gujarāt and fortified by Christian and other religious literature that encouraged similar attitudes, Mahatma Gandhi, the most important leader for independence, appeared to his followers as the quintessence of the Hindu tradition. His austere celibate life was one that the Indian laity had learned to respect implicitly. Gandhi's message reached a wider public than that of any of the earlier reformers.

The Western element in Gandhi's ideology has often been exaggerated. His doctrine of nonviolence can be found in many Hindu sources, although his beliefs were much strengthened by Christian ethical literature and especially by the later writings of Leo Tolstoy. His political technique of passive resistance, satyagraha, also has Indian precedents, although in this he was influenced by Western writers such as the American Henry David Thoreau. The chief innovations in Gandhi's philosophy were his belief in the dignity of manual labour and the equality of women. Precedents for both of these can be found in the writings of some 19th-century reformers, but they have little basis in earlier Indian thought. In many ways Gandhi was a traditionalist. His respect for the cow—which he and other educated Indians rationalized as the representative of Mother Earth—was a factor in the failure of his movement to attract large-scale Muslim support. His insistence on strict vegetarianism and celibacy among his disciples, in keeping with the traditions of Vaishnava ascetic ethics, also caused difficulty among some of his followers. Still, the success of Gandhi represented a political culmination of the movement of popular bhakti begun in South India early in the Christian era.

Gandhi as innovator and traditionalist

The mantle of Mahatma Gandhi fell on Vinoba Bhave, one of his most devoted Mahārāshtrian supporters. For some years after independence Vinoba led a campaign of social service that culminated in the *bhūdān* (land-giving) movement, which persuaded many landowners and wealthy peasants to give fields to landless labourers. This movement had some small success in rural areas, but it gradually lost momentum. Although the memory of Gandhi continues to be revered by most Indians, his policies and principles carry little weight. The great bulk of social service is performed by government agencies rather than by voluntary bodies, whether Gandhian or other.

The religious situation after independence. The increase of nationalism, after the division of India into India and Pakistan in 1947, led to a widening of the gulf between Hindus and Muslims. In the early 1970s it was fashionable in Indian circles to paint the relations of the two religions in earlier centuries as friendly, blaming alien rule for the division of India. In Pakistan the tendency has been to insist that Hindus and Muslims have always been "two nations," even though the Hindus were happy under their Muslim rulers. Neither position is entirely correct. In earlier times there was much mutual influence. But the conservative and rigid moralistic element in Indian Islām gained the upper hand long before British power was consolidated in India, and Islāmic influence on Hinduism remained superficial.

Among the pioneers of nationalism, Tilak glorified the Mahārāshtrian hero Śivājī as the liberator of his country from the alien yoke of the Mughals; and Bankim Chandra Chatterjee's militant ascetics, who pledged to conquer and expel the Muslims, sang a battle hymn that no orthodox Muslim could repeat. British rulers of India did little or nothing to lessen Hindu–Muslim tension, and their policy of separate electorates for the two communities worsened the situation. Many leaders of the Indian National Congress movement, such as Jawaharlal Nehru, carried their Hinduism lightly and favoured a secular approach to politics. The majority, however, followed the lead of Gandhi, whose insistence on Hindu values discouraged Muslims from joining his movement, despite the fact that at his prayer meetings he recited passages from the Qur'ān as well as from Hindu and Christian scriptures. To the right of the Congress politically, the Hindu Mahasabha was equally nationalistic, but its explicitly Hindu nationalism was not opposed to nonviolence in its drive to establish a Hindu state in India.

The transfer of power in 1947 was accompanied by slaughter and pillage of huge proportions. Millions of Hindus left their homes in Pakistan for India, and millions of Muslims migrated in the opposite direction. The tension culminated in the assassination of Gandhi by a Hindu fanatic in January 1948.

India as a secular state

The policy of the Indian government was to establish a secular state, and the successive Congress governments have broadly kept to this policy. The governments of the Indian states, however, have not been so restricted by constitutional niceties. Some state governments have introduced legislation of a specifically Hindu character.

On the other hand, the Congress government has passed legislation more offensive to Hindu traditional prejudices than anything that any British Indian government would have dared to enact. All forms of discrimination against "untouchables" (now usually referred to by euphemisms such as "harijans," or "people of God," instead of the British euphemism "scheduled castes") are forbidden, although it has been impossible to enforce the law in every case. A great blow to conservatism was dealt by legislation in 1955 and 1956 that gave full rights of inheritance to widows and daughters, enforced monogamy, and permitted divorce on quite easy terms. The 1961 law forbidding dowries further undermined traditional Hinduism. Although the dowry has long been a tremendous burden to the parents of daughters, the strength of social custom is such that the law cannot be fully enforced.

The social structure of traditional Hinduism is slowly crumbling in the cities. Intercaste and interreligious marriages are becoming more frequent among the educated, although some aspects of the caste system show remarkable vitality, especially in the matter of appointments and elections. The bonds of the tightly knit Hindu joint family are also weakening, a process helped by legislation and the emancipation of women. The professional priests, who perform rituals for lay people in homes or at temples and sacred sites, complain of the lack of custom, and their numbers are diminishing.

Nevertheless, Hinduism is far from dying. Mythological films, once the most popular form of entertainment, are enjoying a renaissance. Organizations such as the Ramakrishna Mission flourish and expand their activities. New teachers appear from time to time and attract considerable followings. Militant fundamentalist Hindu organizations such as the Society for the Self-Service of the Nation (Rashtriya Svayamsevak Sangh; RSS) are steadily growing. Such movements can be seen as the cause or the result, or both, of persistent outbreaks of communal religious violence involving Hindus and Sikhs in North India, Tamil Hindus and Sri Lankan Buddhists in Sri Lanka, Tamil extremists and moderates in Tamil Nādu, and, still everywhere, Hindus and Muslims.

The adaptability of Hinduism to changing conditions is illustrated by the appearance in the Hindu pantheon of a new divinity, of special utility in an acquisitive society. This is the goddess Santoṣī Mātā, first worshiped widely by women in many cities of Uttar Pradesh and now worshiped throughout India, largely as the result of a popular mythological film about her birth and the origin of her worship. The new goddess was unheard-of a few years ago and has no basis in any Purāṇic myth. Propitiated by comparatively simple and inexpensive rites performed in the home without the intervention of a priest, Santoṣī, it is believed, grants practical and obvious blessings, such as a promotion for a needy, overworked husband, a new radio, or even a refrigerator. News of Santoṣī's blessings is passed from housewife to housewife, and even moderately well-educated women have become her devotees.

The appearance of a new deity: Santoṣī Mātā

On both the intellectual and the popular level, Hinduism is thus in the process of adapting itself to new values and new conditions that have been brought about by mass education and industrialization and is responding to 20th-century challenges.

Hinduism outside India. Since the latter part of the 19th century large colonies of Hindu migrants have been established in East Africa, Malaysia, the islands of the Pacific and the Indian Ocean, and some of the islands of the West Indies. These migrants have taken their religion with them and have adhered to it faithfully for several generations. In recent years they have been aided by Hindu missionaries, chiefly from the Arya Samaj or the Ramakrishna Mission. Since World War II many Hindus have also settled in the United Kingdom. Most of these migrants, however, are comparatively uneducated, and their religion has made little impression on the people among whom they live. They also have made no serious attempts to gain converts. Yet, one of the most striking aspects of contemporary Western culture is its readiness to accept Eastern religious ideas in a way that is unprecedented since the days of the Roman Empire. A recent manifestation of the spread of Indian religious attitudes in the Western world is the Hare Krishna cult, officially known as the International Society for Krishna Consciousness, with its principal office in Los Angeles. This is essentially a bhakti movement, broadly following the precedents of Caitanya. Since its foundation by a Hindu sannyasi, A.C. Bhaktivedanta (Swami Prabhupāda), in 1966, its growth has been surprising, and *saṇkirtana* (devotional singing and dancing) can be seen in the streets of New York City and London, performed by young men and women from Christian or Jewish homes wearing dhotis and saris. These manifestations are part of a process that began in 1784 with the first English translation of a Hindu religious text, Charles Wilkins' version of the *Bhagavadgītā*.

Hinduism is not by nature a proselytizing religion, however, in part because of its inextricable roots in the social system and the land of India. In recent years, many new gurus, such as Bhagwan Shree Rajneesh and Satya Sai Baba, have been successful in making converts in Europe and the United States. The very success of these gurus, however, has produced material profits that many people regard as incompatible with the ascetic attitude appropriate to a Hindu spiritual leader; in some cases, the profits have led to notoriety and even legal prosecution. In addition, the self-proclaimed conversion to questionable forms of "Hinduism" by popular singers and film stars has tended both to increase the glamour and to diminish the respectability of these new forms of Orientalism. That Hinduism is flourishing in India is obvious; that it has made, and can continue to make, a genuine contribution to Western religious thought is undeniable; that the invasion of the gurus is a part of that contribution is highly debatable. (A.L.B./J.A.B.v.B./W.Do.)

Sacred texts

VEDAS

Importance of the Vedas. The Veda, meaning "Knowledge," is a collective term for the sacred scriptures of the Hindus. Since about the 5th century BC, the Veda has been considered to be the creation of neither human nor god; rather, it is regarded as the eternal Truth that was in ancient times directly revealed to or "heard" by gifted and inspired seers (rishis) who transcribed it into the most perfect human language, Sanskrit. Although most of the religion of the Vedic texts, which revolves around rituals of fire sacrifice, has been eclipsed by Hindu doctrines and practices, the absolute authority and sacredness of the Veda remains a central tenet of virtually all Hindu sects and traditions. Even today, as it has been for several millennia, parts of the Veda are memorized and recited as a religious act of great merit.

The components of the Veda. The Veda is the product of the Aryan invaders of the Indian subcontinent and their descendants, although the original inhabitants (disdainfully called *dásyus*, or "slaves," in the Veda) may very well have exerted an influence on the final product. The Veda represents the particular interests of two classes of Aryan society, the priests (Brahmans) and the warrior-kings (Kṣatriyas), who together ruled over the far more numerous peasants (Vaiśyas).

Vedic literature ranges from the Rigveda (Ṛgveda; *c.* 1400 BC) to the Upanishads (Upaniṣads; *c.* 1000–500 BC). This literature provides the sole documentation for all Indian religion before Buddhism and the early texts of classical Hinduism. Because it is the literature of a ruling class, it probably does not represent all the myths and cults of the early Indo-Aryans, let alone those of the non-Aryans.

The great works of Vedic literature

The most important texts are the four collections (Saṃhitās) known as the Veda or Vedas (*i.e.*, "Book[s] of Knowledge"): the Rigveda ("Wisdom of the Verses"), the Yajurveda ("Wisdom of the Sacrificial Formulas"), the Sāmaveda ("Wisdom of the Chants"), and the Athar-

vaveda ("Wisdom of the Atharvan Priests"). Of these, the Rigveda is the oldest.

In the Vedic texts following these earliest compilations, the Brāhmaṇas (discussions of the ritual), Āraṇyakas (books studied in the forest), and Upanishads (secret teachings concerning cosmic equations), the interest in the early Rigvedic gods wanes, and they become little more than accessories to the Vedic rite. Polytheism begins to be replaced by a sacrificial pantheism of Prajāpati ("Lord of Creatures"), who is the All. In the Upanishads Prajāpati merges with the concept of *brahman,* the supreme reality and substance of the universe (not to be confused with the Hindu god Brahmā), replacing any specific personification, thus transforming the mythology into abstract philosophy.

Together, the components of each of the four Vedas—the Saṃhitās, Brāhmaṇas, Āraṇyakas, and Upanishads—constitute the revealed scripture of Hinduism, or the Sruti (*Śruti;* "Heard"). All other works—in which the actual doctrines and practices of Hindus are encoded—are recognized as having been composed by human authors and are thus classed as Smriti (*Smṛti;* "Remembered"). The categorization of Veda, however, is capable of elasticity. First, the Sruti is not exactly closed; Upanishads, for example, have been composed until recent times. Second, the texts categorized as Smriti inevitably claim to be in accord with the authoritative Sruti, and thus worthy of the same respect and sacredness. For Hindus, the Veda is a symbol of unchallenged authority and tradition.

The Rigveda. The religion reflected in the Rigveda is a polytheism mainly concerned with the propitiation of divinities associated with the sky and the atmosphere. Of these, the Indo-European sky father Dyaus was by then little regarded. More important were such gods as Indra, Varuṇa (guardian of the cosmic order), Agni (the sacrificial fire), and Sūrya (the Sun).

The soma sacrifice in the Rigveda

The main ritual activity referred to in the Rigveda is the soma sacrifice. Soma was a hallucinogenic beverage prepared from a now-unknown plant; recently it has been suggested that the plant was a mushroom and that later another plant was substituted for the agaric fungus, which had become difficult to obtain. The Rigveda contains a few clear references to animal sacrifice, which probably became more widespread later. There is some doubt whether the priests formed a separate class of society at the beginning of the Rigvedic period. If they did so, the prevailingly loose boundaries of class made it possible for a man of nonpriestly parentage to become a priest. By the end of the period, however, they had become a separate class of specialists, the Brahmans (Brāhmaṇas), who claimed superiority over all the other social classes, including the Rājanyas (later Kṣatriyas), the warrior-kings.

The Rigveda contains little about birth rituals, but the rites of marriage and disposal of the dead were basically the same as in later Hinduism. Marriage was an indissoluble bond cemented by a lengthy and solemn ritual centring on the domestic hearth. The funeral rites of the rich included cremation, although other funeral forms were also practiced. An interesting reference in one hymn shows that the wife of the dead man lay down beside him on the funeral pyre but was called upon to return to the land of the living before it was lighted. This may have been a survival from an earlier period when the wife was actually cremated with the husband, a custom that was revived in later times.

Among other features of Rigvedic religious life that were important for later generations were the munis. The muni was apparently a sort of shaman (a religious personage having healing and psychic transformation powers), trained in various magic arts and believed to be capable of supernatural feats, such as levitation. He was particularly associated with the god Rudra, a deity connected with mountains and storm and more feared than loved. Rudra developed into the Hindu god Śiva, and his prestige increased steadily. The same is true of Vishnu, a minor solar deity in the Rigveda, who later became one of the most important and popular divinities of Hinduism.

One of the favourite myths of the Aryans was one that attributed the origin of the cosmos to the god Indra, after he had slain the great dragon Vṛtra, a myth very similar to one known in early Mesopotamia. With time, such tales were replaced by more abstract theories that are reflected in several hymns of the late 10th book of the Rigveda. These speculative tendencies were the beginnings of the persistent effort of Indian philosophers to reduce all things to a single basic principle.

Elaborations of text and ritual: the later Vedas. The chronology of later Vedic developments is extremely vague, but it probably encompasses the period from 1000 to 500 BC, which are the dates of the Painted Grayware strata in the archaeological sites of the western Ganges Valley. These excavations reflect a culture still without writing but showing considerable advances in civilization. Nothing, however, has been discovered from sites of this period that throws much light on the religious situation, and historians still must rely on the following texts to describe this phase of the religion.

The other Saṃhitās

The Yajurveda and Sāmaveda. The Yajurveda and Sāmaveda are completely subordinate to the liturgy. The Yajurveda contains the lines, usually in brief prose, with which the executive priest (*adhvaryu*) accompanies his ritual manipulations, addressing the implements he handles and the offering he pours and admonishing other priests to do their invocations. The Sāmaveda is a collection of verses from the Rigveda (and a few new ones) that were chanted with certain fixed melodies.

The Atharvaveda. The Atharvaveda stands apart from other Vedic texts. It contains both hymns and prose passages and is divided into 20 books. Books 1–7 contain magical prayers for precise purposes: spells for a long life, cures, curses, love charms, prayers for prosperity, charms for kingship and Brahmanhood, and expiations for evil committed. They reflect the magical-religious concerns of everyday life and are on a different level than the Rigveda, which glorifies the great gods and their liturgy. Books 8–12 contain similar texts but also include cosmological hymns that continue those of the Rigveda and provide a transition to the more complex speculations of the Upanishads. Books 13–20 celebrate the cosmic principle (book 13) and present marriage prayers (book 14), funeral formulas (book 18), and other magical and ritual formulas. This text is an extremely important source of knowledge of practical religion and magic, particularly where it complements the one-sided picture of the Rigveda. Many rites are also laid down in the "Kauśikasūtra" (manual of the Kauśika family of priests) of the Atharvaveda.

The Brāhmaṇas and Āraṇyakas. Attached to each Saṃhitā was a collection of explanations of the rituals, called a Brāhmaṇa, which often relied on mythology to trace the origins and importance of individual ritual acts. Although they were not manuals or handbooks in the manner of the later Śrauta Sutras, the Brāhmaṇas do contain some detail about the performance and meaning of Vedic sacrificial rituals and are invaluable sources of information about the Vedic religion.

In these texts the sacrifice is the very centre of cosmic processes, all human concerns, and religious desires and goals. It is through the sacrifice that the cosmos continues in its cycles and that human beings obtain the goods of life and a birth in heaven in the next world. The ritual was thought to have such effects on the visible and invisible worlds because of homologies, or connections (*bandhus*), that were said to lie between the components and phases of the ritual and corresponding parts of the universe. The universalization of the dynamics of the ritual into the dynamics of the cosmos was depicted as the sacrifice of the primordial deity, Prajāpati ("Lord of Creatures"), who was perpetually regenerated by the sacrifice.

The lengthy series of rituals of the royal consecration, the *rājasūya,* emphasized royal power and endowed the king with a divine charisma, raising him, at least for the duration of the ceremony, to the status of a god. Typical of this period was the elaborate *aśvamedha,* the horse sacrifice, in which a consecrated horse was freed and allowed to wander at will for a year; it was always followed by the king's troops, who defended it from all attack until it was brought back to the royal capital and sacrificed in a very complicated ritual.

Vedic cosmic-sacrificial speculations continued in the

Āraṇyakas (forest books), which contain materials of two kinds: Brāhmaṇa-like discussions of rites not believed to be suitable for the village (hence the name "forest") and continuing visions of the relationship between sacrifice, universe, and man. The word *brahman*—the creative power of the ritual utterances, which is used to denote the creativeness of the sacrifice and which underlies ritual and therefore cosmic order—is prominent in these texts.

Vedic religion. *Cosmogony and cosmology.* In the Vedic literature there are different but not exclusive accounts of the origin of the universe. The simplest is that the creator built the universe with timber, as a carpenter builds a house. Hence there are many references to gods measuring the different worlds as parts of one edifice, atmosphere upon Earth, heaven upon atmosphere. Creation may be viewed as procreation: the personified Heaven, Dyaus (the word is related to the Greek Zeus), impregnates the Earth goddess, Pṛthivī, with rain, causing crops to grow on her. Quite another myth is recorded in the last (10th) book of the Rigveda: in the "Hymn of the Cosmic Man" ("Puruṣasūkta") it is said that the universe was created out of the parts of the body of a single cosmic man (Puruṣa) when his body was immolated and dismembered at the primordial sacrifice. There the four classes (varnas) of Indian society are referred to: the priest (Brahman) emerging from the mouth, the warrior (Rājanya) from the arms, the peasant (Vaiśya) from the thighs, and the servant (Śūdra) from the legs of the primeval victim. The "Puruṣasūkta" represents the beginning of a new phase, in which the sacrifice became more important and elaborate as cosmological and social philosophies were constructed around it.

Creation

In the same book of the Rigveda, mythology begins to be transformed into philosophy; for example, "in the beginning was the nonexistent, from which the existent arose" (Rigveda 10.72.2). Even the reality of the nonexistent is questioned: "then there was neither the nonexistent nor the existent" (Rigveda 10.129). Such cosmogonic speculations continue, particularly in the older Upanishads. Originally there was nothing at all, or Hunger, which then, to sate itself, creates the world as its food. Alternatively, the creator creates himself in the universe by an act of self-recognition, self-formulation, or self-formation. Or the one creator grows "as big as a man and a woman embracing" (Bṛhadāraṇyaka Upanishad I.4.3) and splits into man and woman, and in various transformations the couple create other creatures. In one of the last stages of this line of thought (Chāndogya Upanishad 6.2), the following account became fundamental to the ontology of the philosophical schools of Vedānta: in the beginning was the Existent, or *brahman,* which through heaven, Earth, and atmosphere (the triadic space) and the three seasons of summer, rains, and harvest (the triadic time) produced the entire universe.

The three worlds of heaven, atmosphere, and Earth

The Vedic texts generally regarded the universe as three layers of "worlds" (*loka*): heaven, atmosphere, and Earth. Heaven is that part of the universe where the sun shines and is correlated with sun, fire, and ether; the atmosphere is that part of the sky between heaven and Earth where the clouds insert themselves in the rainy season and is correlated with water and wind; Earth, a flat disk, like a wheel, is here below as the "holder of treasure" (*vasuṃdharā*) and giver of food. In addition to this tripartite pattern, there is also an ancient notion of duality, in which heaven is masculine and father and Earth is feminine and mother. Later texts present the conception that combinations and permutations of five elements (ether-space [*ākāśa*], wind [*vāyu*], fire [*agni*], water [*āpas*], and earth [*bhūmi*]) formed the universe.

Theology. Generally speaking, Vedic gods share many characteristics: several of them (Indra, Varuṇa, Vishnu) are said to have created the universe, set the Sun in the sky, and propped apart heaven and Earth. All of them are bright and shining, and all are susceptible to human praise. Some major gods were clearly personifications of natural phenomena, and for these deities no clearly delineated divine personalities were perceived.

Principal Vedic deities

The three most frequently invoked gods are Indra, Agni, and Soma. Indra, the foremost god of the Vedic pantheon, is a god of war and rain. Agni (a cognate of the Latin *ignis*) is the deified fire, particularly the fire of sacrifice, and Soma is the intoxicating or hallucinogenic drink of the sacrifice, or the plant from which it is pressed; neither is greatly personified.

The principal focus of Vedic literature is the sacrifice, which in its simplest form can be viewed as a ritualized banquet to which a god is invited to partake of a meal shared by the sacrificer and his priest. The invocations mention, often casually, the past exploits of the deity. The offered meal gives strength to the deity to repeat his feat and to aid the sacrificer.

The myth of Indra killing the dragon Vṛtra has many levels of meaning. Vṛtra prevents the monsoon rains from breaking. Because the monsoon is the greatest single factor in Indian agriculture, the event celebrated in this myth impinges on everyone's life. In the social circles represented in the Rigveda, however, the myth is cast in a warrior mold and the breaking of the monsoon is viewed as a cosmic battle. The entire monsoon complex is involved: Indra is the Lord of the Winds, the gales that accompany the monsoon; his weapons are lightning and thunderbolt, with which he lays Vṛtra low. To accomplish this feat he must be strengthened with soma. Simultaneously, he is the god of war and is invoked to defeat the non-Aryan *dásyu*s, the indigenous peoples referred to in the Vedas. These important concerns—the promptness and abundance of the rains, success in warfare, and the Aryan conquest of the land—all find their focus in Indra.

Because the Vedic gods were not fully anthropomorphic, their functions were subject to various applications and interpretations. Thus Indra, the greatest and most anthropomorphic god of the early Veda, was, in the view of the noble patrons of the Vedic poets, primarily a fighter and a warrior god invoked to bring booty and victory. Agriculturalists and hunters emphasized Indra's fecundity, celebrating his festivals to produce fertility, welfare, and happiness. Indra, however, was essentially a representative of useful force in nature and the cosmos and therefore was the great champion of an ordered and habitable world. His repeated victories over the snake-demon Vṛtra, the representative of obstruction and chaos, resulted in the separation of heaven and Earth (the support of the former and the stabilization of the latter), the rise of the Sun, and the release of the waters: in short, in the organization of the universe.

Although morality is not an issue in Indra's myth, it is in those of the other principal Vedic deities. Central to ancient morality was the notion of rita (*ṛta*), the basic meaning of which appears to have been the truthfulness with which the alliance between humans (and between humans and gods) was observed—a quality necessary to maintain the physical and moral order of the universe. Varuṇa is an older sovereign god, who with Mitra (related to the Persian god Mithra) presides over the observance of the rita. Thus Varuṇa is a judge before whom a mortal may stand guilty, while Indra is a king who may support a mortal king. Typical requests that are made of Varuṇa are for forgiveness, for deliverance from evil committed by oneself or others, and for protection; Indra is prayed to for bounty, for aid against enemies, and for leadership against demons and *dásyu*s.

The notion of rita

Distinct from both is Agni, the fire, who is observed in all his multifarious manifestations: in the sacrificial fire, in lightning, or hidden in the logs from which fire can be drilled. As the fire of sacrifice, he is the mouth of the gods and the carrier of the oblation, the mediator between the human and the divine orders. Agni is above all the good friend of the Aryans and is prayed to to strike down and to burn their enemies and to mediate between gods and men.

Among other Vedic gods, only a few stand out. One is Vishnu, important perhaps more in retrospect than in fact. He is famous for his "three strides," with which he traversed the universe, thus creating and possessing it. In his later mythology this pervasiveness, which invites identification with other gods, remains characteristic. His function as helper to the conqueror-god Indra is important.

Impersonality is increased by the prevalence of pairs and groups of gods. Thus Varuṇa and Mitra are members of the group of Ādityas (sons of Aditi, an old progenitrix),

who generally are celestial gods. They are also combined in the double god Mitra-Varuṇa. Indra and Vishnu are combined as Indra-Vishnu. There is also Rudra, an ambivalent god who is dreaded for his unpredictable attacks but is simultaneously benign insofar as he can restrain his attacks. Although there are many demons (rakshasas), no one god embodies the evil spirit; rather, many gods have their devil within, inspiring fear as well as trust. Among the perpetually beneficent gods are the Aśvins (horsemen), who are helpers and healers and often visit the needy. Almost otiose is the personified heaven, Dyaus, who most often appears literally as the sky, and often as day. As a person, he is coupled with Earth (in the god pair Dyāvā-Pṛthivī) as a father; Earth by herself is more predominantly known as Mother (Mātṛ). Apart from Earth, the other goddess of importance in the text of the Rigveda is Uṣas (Dawn), who brings in the day and thus is said to bring forth the Sun.

Vedic kathenotheism

In the later Vedic period the significance of the Rigvedic gods and their myths began to wane. The peculiar theism of the Rigveda, in which any one of several different gods might be hailed as supreme and attributes of one god could be transferred to another (called kathenotheism by the Vedic scholar F. Max Müller), stressed godhead more than individual gods. In the end this led to a pantheism of Prajāpati, the deified sacrifice or ritualized deity; with his consort Vāc (*i.e.,* the speech of ritual recitation), he is said to have begotten the world.

In the course of the Vedic period Puruṣa fused with the figure Nārāyaṇa ("Scion of Man") and with Prajāpati ("Lord of Creatures"), the patron of procreation in popular belief. In the speculative thought of the ritualists, Prajāpati came to the fore as the creator god and in many respects as the highest divinity, the immortal father even of the gods, whom he transcends, encompasses, and molds into one complex. As the One, the concentrated All, or Totality, Prajāpati was identified with the highest and most general categories. By a process of emanation and self-differentiation (by dividing himself), he created all beings and the universe. After this "creation," Prajāpati became the disintegrated and differentiated All of the phenomenal world and was exhausted. By means of a rite, he then reintegrated himself to prepare for a new phase of creativity. Because the purpose of the sacred act is the restitution of the organic structural norm, which ensures the ordered functioning of the universe, Prajāpati was identified with the rite. Thus, by identifying himself with Prajāpati, a sacrificer may temporarily reintegrate within himself what has been disintegrated, thereby restoring oneness and totality in himself and the universe.

Sin and evil

Ethical and social doctrines. In Vedic times, "sin" (*énas*) or evil (*pāpmán*) was put on a par with illness, enmity, distress, or malediction: it was conceived of as a sort of pollution that could be neutralized by ritual or devices for averting evil. A man might incur "sin" by any incorrect or improper behaviour, especially improper speech, and thus be guilty of *anṛta* (*i.e.,* any infidelity to fact or departure from what is true, real, and constitutes the established order) whether or not he had deliberately committed a crime. Other transgressions included making mistakes in sacrifices and coming into contact with corpses, ritually impure persons, or persons belonging to the lower classes of society. These acts were only rarely considered to be misdeeds against a god or violations of moral principles of divine origin, and the consciousness of guilt was much rarer than the fear of the evil consequences of sin, such as disease or untimely death. Sometimes, however, a god (Agni, the evil-devouring fire, or Varuṇa, the god of order, whose role included punishing and fettering the "sinner") was invoked to forgive the neglect or transgression or to release a man from their concrete results. More usually, however, these results were abrogated by means of purifications, such as the ceremonial use of water, and a variety of expiatory rites.

Importance of ritual purity

To the pure who earned ritual merits, the prospect of a safe "world" (*loka*) or condition was held out. The meticulous effort to purify oneself from every kind of evil also involved the observance of various customs regarding the avoidance of inauspicious occurrences—an endeavour called *śānti*. Ritual purity was the principal concern of the compilers of the manuals of dharma (religious law) that, belonging to the sacred tradition (Smriti; *i.e.,* remembered by human teachers), have contributed much to the special character of Hinduism. According to the authorities on dharma, ritual purity is: the first approach to dharma, the resting place of the Veda (*brahman*), the abode of prosperity (*śrī*), the favourite of the gods, and the means of clearing (soothing) the mind and of seeing (realizing) the atman in the body.

The sacred: nature, man, and God. The contact with the unseen or sacred included humankind's contributions by ritual acts to the maintenance of the universe—of which Vedic thinkers felt themselves an indissoluble part—and to the periodic regeneration, through sacrificial practices, of both the powers for good and the cosmic processes that make earthly life and welfare possible. The Vedic poets were deeply convinced that the world is an organized cosmos governed by order and truth (rita) and that it is always in danger of being damaged or destroyed by the powers of chaos (*asat*). This conviction found mythological expression in the continual conflict between gods (devas) and demoniac antigods (asuras).

Gods were conceived as presiding over certain provinces of the universe or as responsible for important cosmic or social phenomena. Their deeds are timeless and exemplary presentations of mythic events replete with power and universal, eternal significance. To reproduce themselves in time and thus retain their vitality and efficacy, mythical events need to be repeated—that is, celebrated and confirmed by means of the spoken word and ritual acts.

Vedic and Brahmanic rites. Vedic religion is primarily a liturgy differentiated in various types of ritual designed for almost any conceivable purpose. These rites are described in the texts in minute detail; theoretically, no operation, no gesture, no formula is meaningless or left to an officiant's discretion. On the basis of a complicated speculative system, all are explained and shown to be effective in the Brāhmaṇas. The often complicated ritual technique was devised mainly to safeguard human life and survival, to enable people to face the many risks and dangers of existence, to thwart the designs of human and superhuman enemies that cannot be counteracted by ordinary means, to control the unseen powers, and to establish and maintain beneficial relations with the supramundane sacred order. Belief in the efficacy of the rites is the natural consequence of the belief that all things and events are connected with or participate in one another.

Cosmic role of sacred places

Hence it is also believed that a close correspondence exists between a sacred place—such as the sacrificial place of many Vedic rites, a place of pilgrimage, or a consecrated area (*maṇḍala,* "circle")—and a province of the universe or even the universe itself. These places represent, within the reach of the officiants, the universe or as much of it as is relevant. In such places, direct communication with other cosmic regions (heaven or underworld) is possible because they are said to be at the point of contact between this world and the "pillar of the universe," "the navel of the earth." The sacred place is (by virtue of a system of connections) identical with the universe in its various states of emanation from, reabsorption into, integration with, and disintegration from the sacred. This idea has as its corollary the possibility of ritually enacting the cosmic drama and, thus, of influencing, through the same system of connections, those events in the cosmos that continuously affect human weal and woe.

The Vedic ritual system is organized into three main forms. The simplest, and hierarchically inferior, type of Vedic ritualism is the *gṛhya,* or domestic ritual, in which the householder himself offers modest oblations into the one sacred household fire. The more ambitious, wealthy, and powerful married householder sets three or five fires and, with the help of professional officiants, engages in the more complex *śrauta* sacrifices. These require oblations of vegetable substances and, in some instances, of parts of ritually killed animals (mostly goats, but also sheep, cows, horses, and perhaps at one time human beings as well). Finally, at the highest level of Vedic ritualism are the sacrifices of soma, which can continue for days or even years

and whose intricacies and complexities are truly stunning.

In the major *śrauta* rites, requiring three fires and 16 priests or more, "the man who knows"—he who has an insight into the correspondences (*bándhu*) between the mundane and cosmic phenomena and the eternal transcendent reality beyond them and who knows the meaning of the ritual words and acts—may set great cosmic processes in motion for the sake of human interests. In these rites, Brahman officiants repeat the mythic drama for the benefit of their patron, the "sacrificer," who temporarily becomes its centre and realizes through ritual symbolism his identity with the universe. Whatever magical elements may be involved in this ritual technique, its aim in establishing an efficacious contact with a transcendental order that is the source of all life and power is based on an essentially religious conception. Such officiants are firmly convinced of the efficacy of their rites: "the sun would not rise, were he [the officiant] not to make that offering; this is why he performs it" (Śatapatha Brāhmaṇa 2.3.1.5). The oblations should not be used to propitiate the gods or to thank them for favours bestowed, since the efficacy of the rites, some of which are still occasionally performed, does not depend on the will of the gods.

The Upanishads. With the last component of the Veda, the mystically oriented and originally esoteric texts known as the Upanishads, Vedic ritualism and the doctrine of the interconnectedness of separate phenomena was superseded by a new emphasis on knowledge alone—primarily knowledge of the ultimate identity of all phenomena, which merely appeared to be separate. The phase of Indian religious life roughly between 700 and 500 BC was the period of the beginnings of philosophy and mysticism marked by the Upanishads ("Sittings Near a Teacher"). Historically, the most important of these are the two oldest, the Bṛhadāraṇyaka ("Great Forest Text") and the Chāndogya (pertaining to the Chandogas, a class of priests who intone hymns at sacrifices), both of which are compilations that record the traditions of sages (rishis) of the period, notably Yājñavalkya, who was a pioneer of new religious ideas.

The primary motive of the Upanishads is a desire for mystical knowledge that would ensure freedom from "redeath." Throughout the later Vedic period, the idea that the world of heaven was not the end—and that even in heaven death was inevitable—had been growing. For Vedic thinkers, the fear of the impermanence of religious merit and its loss in the hereafter, as well as the fear-provoking anticipation of the transience of any form of existence after death, culminating in the much-feared repeated death (*punarmṛtyu*), assumed the character of an obsession. The means of escaping and conquering death and of attaining integral life devised in the Brāhmaṇas were of a ritual nature, but in one of the oldest Upanishads, the Bṛhadāraṇyaka (*c.* 10th–5th century BC), more emphasis was placed on the knowledge of the cosmic connection underlying ritual. When the doctrine of the identity of atman (the Self) and *brahman* was established in the Upanishads, the true knowledge of the Self and the realization of this identity was (by those sages who were inclined to meditative thought) substituted for the ritual method.

Development of atman-*brahman* doctrine from ritual speculation

In the following centuries, the main theories connected with the divine essence underlying the world were harmonized and synthetically combined, and the tendency was to extol one god as the supreme Lord and Originator (Īśvara), who is at the same time Puruṣa and Prajāpati and *brahman* and the inner Self (atman) of all beings. For those who worshiped him, he became the goal of identificatory meditation, which leads to complete cessation of phenomenal existence and becomes the refuge of those who seek eternal peace.

The period during which the Upanishads were composed was one of much social, political, and economic upheaval. Rural tribal society was disappearing, and the adjustments of the people to urban living under a monarchy probably provoked many psychological and religious responses. During this period many groups of mystics, world-renouncers, and forest-dwellers appeared in India, and these included the authors of the Upanishads. Among the more important practices and doctrines of these world-renouncers were asceticism and the concept of rebirth or transmigration.

The Rigveda shows few examples of asceticism, except among the munis (shamans). The Atharvaveda describes another class of religious adepts, or specialists, the *vrātya*s, particularly associated with the region of Magadha (west central Bihār). The *vrātya* was a wandering hierophant (one who manifested the Holy) who remained outside the regular system of Vedic religion. He traveled from place to place in a bullock cart with an apprentice and with a woman who appears to have been used for ritual prostitution. Flagellation and other forms of self-mortification seem to have been part of his routine. Efforts were made by the orthodox to bring the *vrātya*s into the Vedic system by special rituals of conversion, and it may be that these people helped to introduce non-Aryan beliefs and practices into Vedic religion. At the same time, the more complex sacrifices of the later Vedic period demanded purificatory rituals, such as fasting and vigil, as part of the preparations for the ceremony. Thus there was a growing tendency toward the mortification of the flesh.

The origin and the development of the belief in the transmigration of souls are very obscure. A few passages suggest that this doctrine was known even in the days of the Rigveda, but it was first clearly propounded in the earliest Upanishad—the Bṛhadāraṇyaka. There it is stated that normally the soul returns to Earth and is reborn in human or animal form. This doctrine of samsara (reincarnation) is attributed to the sage Uddālaka Āruṇi, who is said to have learned it from a Kṣatriya chief. In the same text, the doctrine of karma (actions), according to which the soul achieves a happy or unhappy rebirth according to its works in the previous life, also occurs for the first time, attributed to Yājñavalkya. Both doctrines appear to have been new and strange ones, circulating among small groups of ascetics who were disinclined to make them public, perhaps for fear of the orthodox priests. These doctrines must have spread rapidly, for in the later Upanishads and in the earliest Buddhist and Jain scriptures they are common knowledge.

SUTRAS, SHASTRAS, AND SMRITIS

The Vedāṅgas. Toward the end of the Vedic period, and more or less simultaneously with the production of the principal Upanishads, concise, technical, and usually aphoristic texts were composed about various subjects relating to the proper and timely performance of the Vedic sacrificial rituals. These were eventually labeled as Vedāṅgas ("Studies Accessory to the Veda").

Early scholarly disciplines

The intense preoccupation with the liturgy gave rise to scholarly disciplines that were part of the Vedic erudition. There were six such fields: (1) *śikṣā* (instruction), which explains the proper articulation and pronunciation of the Vedic texts. Different branches had different ways of pronouncing the texts, and these variations were recorded in *prātiśākhya*s (literally, "Instructions for the *śākhā*s"—branches), four of which are extant; (2) *chandas* (metre), of which there remains only one late representative; (3) *vyākaraṇa* (analysis and derivation), in which the language is grammatically described—Pāṇini's famous grammar (*c.* 400 BC) and the *prātiśākhya*s are the oldest examples of this discipline; (4) *nirukta* (lexicon), which discusses and gives meanings for difficult words, represented by the *Nirukta* of Yāska (*c.* 600 BC); (5) *jyotiṣa* (luminaries), a system of astronomy and astrology used to determine the right times for rituals; and (6) *kalpa* (mode of performance), which studies the correct ways of performing the ritual.

Of special importance are the texts constituting the Kalpa Sutras (collections of aphorisms on the mode of ritual performance). The composition of these texts was begun around 600 BC by Brahmans belonging to the ritual schools (*śākhā*s), each of which was attached to a particular recension of one of the four Vedas. A complete Kalpa Sutra contains four principal components: (1) a Śrauta Sutra, which establishes the rules for performing the more complex rituals of the Vedic repertoire; (2) a Śulba Sutra, which shows how to make the geometric calculations necessary for the proper construction of the ritual arena; (3) a Gṛhya Sutra, which explains the rules for performing the

domestic rites, including the life-cycle rituals (called the samskaras); and (4) a Dharma Sutra, which provides the rules for the conduct of life.

Society was ritually stratified in the four classes, each of which had its own dharma (law), or eternal norm of conduct. The ideal life was constructed through sacraments in the course of numerous ceremonies, performed by the upper classes, that carried the individual from conception to cremation in a series of complex rites. The Gṛhya Sutras show that in the popular religion of the time there were many minor divinities who are rarely mentioned in the literature of the large-scale sacrifices but who were probably far more influential on the lives of most people than were the great gods of Vedism.

Dharma Sutra and Dharma Shastra. Among the texts inspired by the Veda are the Dharma Sutras, or manuals on dharma, which contain rules of conduct and rites as they were practiced in a number of branches of the Vedic schools. Their principal contents address the duties of people at various stages of life or ashramas (studenthood, householdership, retirement, and asceticism); dietary regulations; offenses and expiations; and the rights and duties of kings. They also discuss purification rites, funerary ceremonies, forms of hospitality, and daily oblations. Finally, they even mention juridical matters. The more important of these texts are the sutras of Gautama, Baudhāyana, and Āpastamba. Although the direct relationship is not clear, the contents of these works were further elaborated in the more systematic Dharma Shastras, which in turn became the basis of Hindu law.

Hindu law

First among them stands the Dharma Shastra of Manu, also known as the *Manu-smṛti* ("Tradition of Manu"; *c.* AD 200), with 2,694 stanzas divided into 12 chapters. It deals with various topics such as cosmogony, definition of dharma, the sacraments, initiation and Vedic study, the eight forms of marriage, hospitality and funerary rites, dietary laws, pollution and purification, rules for women and wives, royal law, 18 categories of juridical matters, and finally more religious matters, including donations, rites of reparation, the doctrine of karma, the soul, and punishment in hell. Law in the juridical sense is thus completely embedded in religious law and practice. The framework is provided by the model of the four-class society. The influence of the Dharma Shastra of Manu has been enormous, as it provided Hindu society with its practical morality. For large parts of the Indian subcontinent, Manu's text—mediated by its commentaries, notably that of Medhātithi (9th century)—has been the law.

Second only to Manu is the Dharma Shastra of Yājñavalkya; its 1,013 stanzas are distributed under the three headings of good conduct, law, and expiation. Its commentary, *Mitākṣarā* of Vijñāneśvara (11th century), has extended its influence.

Smriti texts. The shastras are a part of the Smriti ("Remembered," or traditional) literature which, like the sutra literature that preceded it, stresses the religious merit of gifts to Brahmans. Because kings often transferred the revenues of villages or groups of villages to Brahmans, either singly or in corporate groups, the status and wealth of the priestly class rose steadily. In the *agrahāras,* as the settlements of Brahmans were called, they were encouraged to devote themselves to the study of the Vedas and the subsidiary studies associated with them; but many Brahmans also developed the sciences of the period, such as mathematics, astronomy, and medicine, while others cultivated literature.

Family law and the Smriti texts

The Smriti texts are binding to this day on orthodox Hindus, and until quite recently Hindu family law was based on them. Although there is evidence of divorce in early Indian history, by the Gupta period marriage was solemnized by lengthy sacred rites and was virtually indissoluble. Intercaste marriage was becoming rarer and more difficult, and child marriage and the rite of suttee were already in existence, although less frequent than they later became. One of the earliest definite records of a widow burning herself on her husband's pyre is found in an inscription from Eran, Madhya Pradesh, dated 510, but the custom had been followed sporadically long before this. From the 6th century AD onward, such occurrences became frequent in certain parts of India, particularly in Rājasthān.

EPICS AND PURĀṆAS

During the centuries immediately preceding and following the beginning of the Christian era, the recension of the two great Sanskrit epics, the *Mahābhārata* and the *Rāmāyaṇa,* took shape out of existing material such as heroic epic stories, mythology, philosophy, and above all the discussion of the problem of dharma. Much of the material of which the epics are composed dates far back into the Vedic period, while the rest continued to be added until well into the medieval period. It is conventional, however, to date the recension of the Sanskrit texts to the period from 300 BC to AD 300 for the *Mahābhārata* and to the period from 200 BC to AD 200 for the *Rāmāyaṇa.*

The Mahābhārata. The *Mahābhārata* ("Great Epic of the Bhārata Dynasty"), a text of some 100,000 verses attributed to the sage Vyāsa, was preserved both orally and in manuscript form for centuries. The central plot concerns a great battle between the five sons of Pāṇḍu, called the Pāṇḍavas (Arjuna, Yudhiṣṭhira, Bhīma, and the twins Nakula and Sahadeva), and the sons of Pāṇḍu's brother Dhṛtarāṣṭra. The battle eventually leads to the destruction of the entire race, save for one survivor who continues the dynasty. As each of the heroes is the son of a god (Indra, Dharma, Vāyu, and the Aśvins, respectively), the epic is deeply infused with religious implications. There are, moreover, many passages in which dharma is systematically treated, so that Hindus regard the *Mahābhārata* as one of the Dharma Shastras. Religious practice takes the form of Vedic ritual on official occasions, pilgrimage, and, to some extent, adoration of gods. Apart from the *Bhagavadgītā* (part of book 6) much of the didactic material is found in the Book of the Forest (book 3), in which sages teach the exiled heroes, and in the Book of Peace (book 12), in which the wise Bhīṣma expounds on religious and moral matters.

The Vedic gods have lost importance and survive as figures of folklore. Prajāpati of the Upanishads is popularly personified as the god Brahmā, who creates all classes of beings and dispenses boons. Of far greater importance is Krishna. In the epic he is a hero, a leader of his people, and an active helper of his friends. His biography as it is known later is not worked out; still, the text is the source of early Krishnaism. Not everywhere, and certainly not by everyone, is Krishna considered a god, and even as god his stature is superhuman rather than divine. He is occasionally, but not significantly, identified with Vishnu. Later, as one of the most important of the incarnations of Vishnu, Krishna undergoes a complex development as an incarnate god. In the *Mahābhārata* he is primarily a hero, a chieftain of a tribe, and an ally of the Pāṇḍavas, the heroes of the *Mahābhārata.* He accomplishes heroic feats with the Pāṇḍava prince Arjuna. Typically he helps the Pāṇḍava brothers to settle in their kingdom and, when the kingdom is taken from them, to regain it. In the process he emerges as a great teacher who reveals the *Bhagavadgītā,* the most important religious text of Hinduism. In the further development of the Krishna myth, this dharmic aspect recedes and makes way for an idyllic myth about Krishna's boyhood, when he plays with and loves young cowherd women (*gopīs*) in the village while hiding from an uncle who threatens to kill him. The influence of this theme on art has been profound. But there is a shadow side to this idyll. Even in the *Mahābhārata,* where it is often said that Krishna becomes incarnate in order to sustain dharma when it wanes and to combat adharma (forces contrary to dharma), he himself commits a number of deeds in direct violation of the warrior ethic and is indirectly responsible for the destruction of his entire family. This adharmic shadow is also cast in the Purāṇic idyll because the *gopīs* that he woos are the wives of other men.

Krishna in the *Mahābhārata*

Far remoter than the instantly accessible Krishna is Śiva, who also is hailed as the supreme god in several myths recounted of him, notably the Story of the Five Indras, Arjuna's battle with him, and his destruction of the sacrifice of Daksha. The epic is rich in information about sacred places, and it is clear that making pilgrimages and

bathing in sacred rivers constituted an important part of religious life. Occasionally these sacred places are associated with sanctuaries of gods. More frequent are accounts of mythical events concerning the particular place and enriching its sanctity. Numerous descriptions of pilgrimages (*tīrthayātrā*) give the authors opportunities to detail local myths and legends. In addition to these, countless edifying stories shed light on the religious and moral concerns of the age. Almost divine are the towering ascetics capable of fantastic feats, whose benevolence is sought and whose curses are feared.

The Rāmāyaṇa. The classical narrative of Rāma is recounted in the Sanskrit epic the *Rāmāyaṇa* by the sage Vālmīki, who is the traditional author of the epic. Rāma is deprived of the kingdom to which he is heir and is exiled to the forest with his wife Sītā and his brother Lakṣmaṇa. While there, Sītā is abducted by Rāvaṇa, the demon king of Laṅkā. In their search for Sītā, the brothers ally themselves with a monkey king whose general, Hanumān (who later became a monkey deity), finds Sītā in Laṅkā. In a cosmic battle, Rāvaṇa is defeated and Sītā rescued. When Rāma is restored to his kingdom, the populace casts doubt on Sītā's chastity while a captive. To reassure them, Rāma banishes Sītā to a hermitage, where she bears him two sons and eventually dies by reentering the earth from which she had been born. Rāma's reign becomes the prototype of the harmonious and just kingdom, to which all kings should aspire. Rāma and Sītā set the ideal of conjugal love; Rāma's relationship to his father is the ideal of filial love; and Rāma and Lakṣmaṇa represent perfect fraternal love. Everything in the myth is designed for harmony, which after being disrupted is at last regained.

Rāma and Sītā

In all but its oldest form, the *Rāmāyaṇa* identifies Rāma with Vishnu as another incarnation and remains the principal source for Rāmaism (worship of Rāma). Though not as long as the *Mahābhārata,* the text contains a great deal of comparable religious material in the form of myths, stories of great sages, and accounts of exemplary human behaviour.

Rāma also has a shadow side. His killing of the monkey king Vālin (or Bālin) in violation of all rules of combat and his banishment of the innocent Sītā are troublesome to subsequent tradition. These problems of the "subtlety" of dharma and the inevitability of its violation, central themes in both epics, remained the locus of philosophical argument throughout Indian history. Apart from their influence as Sanskrit texts, the *Mahābhārata* and the *Rāmāyaṇa* have made an impact in southern and Southeast Asia, where their stories have been continually retold in vernacular and oral versions, and their influence on Indian and Southeast Asian art has been profound. Even today, the epic stories and tales are part of the early education of all Hindus; a continuous reading of the *Rāmāyaṇa* is an act of great merit, and a popular enactment of one version is an annual event across northern India.

The Bhagavadgītā. The *Bhagavadgītā* ("Song of the Lord") is the most influential Indian religious text, although it is not strictly Sruti, or revelation. It is a brief text, 700 verses divided into 18 chapters, in quasi-dialogue form. When the opposing parties in the *Mahābhārata* war stand ready to begin battle, Arjuna, the hero of the favoured party, despairs at the thought of having to kill his kinsmen and lays down his arms. Krishna, his charioteer, friend, and adviser, thereupon argues against Arjuna's failure to do his duty as a noble. The argument soon becomes elevated into a general discourse on religious and philosophical matters. The text is typical of Hinduism in that it is able to reconcile different viewpoints, however incompatible they seem to be, and yet emerge with an undeniable character of its own. Three different ways of releasing the self from transmigration are set forth. There is the discipline of action (karma-yoga): against the views held by Buddhism, Jainism, and Sāṃkhya philosophy, which hold that all acts bind and that therefore abstention from action is a precondition of release, Krishna argues that it is not the acts that bind but the selfish intentions with which they are performed. He argues for a self-discipline in which a person does his duties according to the dictates of prescribed tasks (dharma), but without any self-interest in the personal consequences of the acts. On the other hand, he does not deny the relevance of the discipline of knowledge (jnana-yoga), in which one seeks release in a yogic (ascetic) course of withdrawal and concentration. Then the tone changes and becomes intensely religious: Krishna reveals himself as the Supreme God and grants Arjuna a vision of himself. The third, and perhaps superior, way of release is through a discipline of devotion to God (bhakti-yoga) in which the self humbly worships the loving God and in release hopes not so much for personal liberation from transmigration but for an eternal vision of God. In response to this devotion, God will extend his grace to his votaries, enabling them to overcome the bonds of this world.

Theologies of the *Bhagavadgītā*

The *Bhagavadgītā* is not a systematic theological treatise, and it combines many different elements from Sāṃkhya and Vedānta philosophy. In matters of religion, its important contribution was the new emphasis placed on devotion, which has since remained a central path in Hinduism. In addition, the popular theism evidenced elsewhere in the *Mahābhārata* and the transcendentalism of the Upanishads converge, and a God of personal characteristics is identified with the *brahman* of the Vedic tradition. In its three disciplines the *Bhagavadgītā* gives a typology of the three dominant trends of Indian religion: dharma-based Brahmanism, enlightenment-based asceticism, and devotion-based theism.

The influence of the *Bhagavadgītā* has been profound. It was a popular text, open to all who would listen, and it was fundamental for all later Hinduism. Vedānta philosophy recognizes it, with the Upanishads and the *Brahmasūtra*s (brief doctrinal rules concerning *brahman*), as the third authoritative text, so that all philosophers wrote commentaries on it. Even in the 20th century, as is evident from the lives of such diverse personalities as the Indian freedom fighters Tilak and Gandhi, who acknowledged its influence, it has continued to shape the attitudes of Hindus.

The *Bhagavadgītā,* by demanding that God's worshipers fulfill their duties—"better one's own duty ill-done than another's well-performed" (3.35)—and observe the rules of moral conduct, bridged the chasm between ascetic disciplines and the search for emancipation, on the one hand, and the exigencies of daily life, on the other. For those who must lead a normal life in this world, the *Bhagavadgītā* gave a moral code and a prospect of final liberation. Thus, the work founded, on the basis of the Vaishnava tradition, what may be called a social ethic. Because God is in all beings as their physical and psychical substratum—and as he exists collectively in human society—the wise should not see any difference between their fellow creatures and should love God in them equally. Like God himself, the devotee should be impartial—the same to friend as to foe. The serious endeavour of realizing God's presence in human beings requires humility and a complete unconsciousness of oneself as a corollary of the consciousness of the Presence. It demands the selfless dedication of all actions, duties, and ceremonies to the Lord and obliges a person to promote both individual and social uplift and welfare. Yet, by emphasizing that all humans have not only different propensities for each of the three disciplines of release but also different responsibilities arising out of their births in different castes, the *Bhagavadgītā* also provided a powerful justification for the caste system.

The Purāṇas. The period of the Guptas saw the production of the first of the series (traditionally 18) of often voluminous texts that treat in encyclopaedic manner the myths, legends, and genealogies of gods, heroes, and saints. The usual list of the *Purāṇa*s is as follows: the *Brāhma-, Brāhmāṇḍa-, Brahmavaivarta-, Mārkaṇḍeya-, Bhaviṣya-,* and *Vāmana-Purāṇa*s; the *Viṣṇu-, Bhāgavata-, Nāradīya-, Gāruḍa-, Pādma-,* and *Vārāha-Purāṇa*s; and the *Śiva-, Liṅga-, Skanda-, Agni-* (or *Vāyu-*), *Mātsya-,* and *Kūrma-Purāṇa*s. Many deal with the same or similar materials.

Popular Sanskrit texts

With the epics, with which they are closely linked in origin, the *Purāṇa*s became the scriptures of the common people; they were available to everybody, including women and members of the lowest order of society (Śūdras), and were not, like the Vedas, restricted to initiated

men of the three higher orders. The origin of much of their contents may be non-Brahmanic, but they were accepted and adapted by the Brahmans, who thus brought new elements into their orthodox religion.

At first sight the discontinuity between Vedic and Purāṇic mythology appears to be so sharp that they might be considered as being of altogether different traditions. Yet it soon becomes clear that they are in part continuous and that what appears to be discrepancy is merely a difference between the liturgical emphasis of the Vedas and the more eclectic genres of the epics and *Purāṇa*s. For example, the great god of the Rigveda is Indra, the god of war and monsoon, prototype of the warrior; but for the population as a whole he was more important as the rain god than the war god, and it is as such that he survives in early Purāṇic mythology. Little is learned in the Veda of goddesses, yet they rose steadily in recognition in Purāṇic mythology.

Although in the *Purāṇa*s some of the Vedic gods have an afterlife in which their importance is reduced, other gods, previously of less official significance, arise. The two principal gods of Purāṇic Hinduism are Vishnu and Rudra-Śiva. Both are known in the Vedas, though they play only minor roles: Vishnu is the strider who, with his three strides, established the three worlds (heaven, atmosphere, and Earth) and thus is present in all three orders; and Rudra-Śiva is a mysterious god who must be propitiated.

Purāṇic literature documents the stages of the rise of the two gods as they eventually attract to themselves the identities of other popular gods and heroes: Vishnu assumes the powers of those gods who protect the world and its order, Śiva the powers that are outside and beyond Vishnu's. To these two is often added Brahmā, creator of the world and teacher of the gods. Although still a cosmic figure, Brahmā appears in the *Purāṇa*s primarily to appease overpowerful sages and demons by granting them boons.

In the Purāṇic literature of AD 500 to 1000, sectarianism creeps into mythology, and one god is extolled above the others. Of prime interest are cosmology, myths of the great ascetics (who in some respects eclipse the old gods), and myths of sacred places, usually rivers and fords, whose powers to reward the pilgrim are often cited and related to local legends.

Mythologies of creation and doomsday

Cosmogony. Purāṇic cosmogony greatly expands upon the already complex cosmogonies of the Brāhmaṇas, Upanishads, and epics. According to one of many versions of the story of the origin of the universe, in the beginning the god Nārāyāṇa (identified with Vishnu) floated on the snake Ananta ("Endless") on the primeval waters. From his navel grew a lotus, in which the god Brahmā was born reciting the four Vedas with his four mouths and creating the "Egg of Brahmā," which contains all the worlds. There are numerous other accounts that refer to other demiurges, or creators, like Manu (the primordial ancestor of humankind).

Although the Vedas do not seem to conceive of an end to the world, Purāṇic cosmogony accounts for the periodic destruction of the world at the close of an eon, when the Fire of Time will put an end to the universe. Elsewhere the destruction is specifically attributed to the god Śiva, who dances the tandava dance of doomsday and destroys the world. Yet this end is not an absolute end but a temporary suspension (pralaya), after which creation begins again in the same fashion.

Cosmology. The *Purāṇa* texts present an elaborate mythical cosmography. The old tripartite universe persists, but it is modified. There are three levels—heaven, Earth, and the netherworld—but the first and last are further subdivided into vertical layers. Earth consists of seven circular continents, the central one surrounded by the salty ocean and each of the other concentric continents by oceans of other liquids. In the centre of the central mainland stands the cosmic mountain Meru; the southernmost portion of this mainland is Bhāratavarṣa, the old name for India. Above Earth there are seven layers in heaven, at the summit of which is the world of Brahmā (*brahma-loka*); there are also seven layers below Earth, the location of hells inhabited by serpents and demons of various kinds.

Myths of time and eternity. The oldest texts speak little of time and eternity. It is taken for granted that the gods, though born, are immortal; they are called "sons of Immortality." In the Atharvaveda, Time appears personified as creator and ruler of everything. In the Brāhmaṇas and later Vedic texts there are repeated esoteric speculations concerning the year, which is the unit of creation and thus is identified with the creative and regenerative sacrifice and with Prajāpati ("Lord of Creatures"), the god of the sacrifice. Time is an endless repetition of the year, and thus of creation; this is the starting point of later notions of repeated creations.

The four yugas of the world

Purāṇic myths develop around the notion of yuga (world age), of which there are four. These four yugas, Kṛta, Tretā, Dvāpara, and Kali—they are named after the four throws, from best to worst, in a dice game—constitute a *mahāyuga* (large yuga), and, like the comparable ages of the world depicted by the Greek poet Hesiod, are periods of increasing deterioration. Time itself also deteriorates, for the ages are successively shorter. Each yuga is preceded by an intermediate "dawn" and "dusk." The Kṛta Yuga lasts 4,000 years, with a dawn and dusk of 400 years each, or a total of 4,800 years; Tretā a total of 3,600 years; Dvāpara 2,400 years; and Kali (the current one), 1,200 years. A *mahāyuga* thus lasts 12,000 years and observes the usual coefficient of 12, derived from the 12-month year, the unit of creation. These years are "years of the gods," each lasting 360 human years, 360 being the days in a year. Two thousand *mahāyuga*s form one kalpa (eon), which is itself but one day in the life of Brahmā, whose full life lasts 100 years; the present is the midpoint of his life. Each kalpa is followed by an equally long period of abeyance (pralaya), in which the universe is asleep. Seemingly the universe will come to an end at the end of Brahmā's life, but Brahmās too are innumerable, and a new universe is reborn with each new Brahmā.

Another myth lays particular stress on the destructive aspect of time. Everything dies in time: "Time ripens the creatures, Time rots them" (*Mahābhārata* 1.1.188). "Time" (*kāla*) is thus another name for the god of death, Yama. The name is associated especially with Śiva in his destructive aspect as Mahākāla and is extended to his consort, who may be known as the goddess Kālī or Mahākālī. On a mythological level the speculations on time reflect the doctrine of the eternal return in the philosophy of transmigration. The universe returns just as, after death, a soul returns to be born again. In the oldest description of the process (Chāndogya Upanishad 5.3.1.–5.3.10), the account is still mythic, but with tendencies to naturalism. The soul on departing may go either of two ways: the Way of the Gods, which brings it through days, bright fortnights, the half year of the northern course of the Sun, to the full year, and eventually to *brahman;* or the Way of the Ancestors, through nights, dark fortnights, the half year of the southern course of the Sun, and, failing to reach the full year, eventually back to Earth clinging to raindrops. If the soul happens to light on a plant that is subsequently eaten by a man, the man may impregnate a woman and thus the soul is reborn. Once more the significance of the year as a symbol of complete time is clear.

Myths of the gods. According to the epic *Mahābhārata* (1.1.39), there are 33,333 Hindu deities. In other, later sources, that number is multiplied a thousandfold. Usually, however, the gods are referred to as "The Thirty-Three."

The tendency toward pantheism increased in Purāṇic Hinduism and led to a kind of theism that exalted several supreme gods who were not prominently represented in the Vedic corpus, while many of the Vedic gods disappeared or were greatly diminished in stature. New patterns became apparent: the notion of rita, the basis of the conception of cosmic order, was reshaped into that of dharma, or the religious-social tasks and obligations of humans in society that maintain order in the universe. There also was a broader vision of the universe and the place of divinity.

Three principal moments are envisioned in the life of the cosmos: creation, maintenance, and destruction. Important myths about the gods are tied to these moments. Traditionally, Brahmā is the creator, emanating the universe and simultaneously promulgating the four Vedas from his four mouths. The conception of time as almost

endlessly repeating itself in kalpas detracts, however, from the uniqueness of the first creation, and Brahmā becomes little more than a demiurge. Far more attention is given to the maintenance and to the destruction of the universe.

Maintenance and destruction are symptomatic of order and disorder, and order and disorder in turn are closely associated with society and the realm outside society. The god Vishnu, who became the god of maintenance, is thus also the social god par excellence, while Śiva, partly established as the agent of destruction, is in many respects an asocial god. Vishnu is the saviour from lawlessness, destroyer of those who threaten the good order, and king of the harmonious realm. Śiva represents untamed wildness; he is the lone hunter and dancer, the yogi (the accomplished practitioner of yoga) withdrawn from society, and the ash-covered ascetic. The distinction between the gods is not between good and evil but rather between two ways in which the divine manifests itself in this world—as both benevolent and fearful, both harmonious and disharmonious.

Bhāgavata-Purāṇa. South Indian devotionalism produced many works in Sanskrit; the most important was the *Bhāgavata-Purāṇa,* which soon became known throughout India. Its 10th book is devoted entirely to Krishna, and there, for the first time, his adventures as a lovable child and as a youth are recounted in great detail. The *Bhāgavata-Purāṇa* may have been written in the 10th century and is certainly a product of the Dravidian south. The doctrine of the avatars of Vishnu was by now in full force, and the *Bhāgavata* recognizes 22 of them.

While all *Purāṇa*s have exerted influence on Hinduism—and are in turn reflections of trends in Hinduism—none can compare in popularity with the *Bhāgavata-Purāṇa* ("The *Purāṇa* of the Devotees of the Blessed Lord Krishna"), the *Purāṇa* of the god Krishna par excellence. It differs from the other *Purāṇa*s in that it is planned as a unit and that far greater care is taken with both metre and style. Its nearly 18,000 stanzas are divided into 12 books. The most popular part of the *Purāṇa* is the description of the life of Krishna, for which it has since remained the principal authority. In this work far greater emphasis than in other texts is placed on the youth of Krishna: the threats against his life by the tyrant Kaṃsa, his flight and life among the cowherds at Gokula, and especially his adventures and pranks with the cowherd girls. This treatment has remained classic, and the popularity of the text has led to the survival of many manuscripts, some beautifully illustrated. Much of medieval Indian painting and an enormous amount of vernacular literature draw upon the *Bhāgavata-Purāṇa* for their themes.

The childhood of Krishna

The *Bhāgavata-Purāṇa* teaches a quite representative Vaishnava theology: God is transcendent and beyond human understanding; he is the universal causality, creator and substratum; he is time and the bearer of all possibilities that are susceptible of actualization; through his incomprehensible creative ability (maya) or specific power (*ātmaśakti*) he expands himself into the universe, which he pervades and which is his outward appearance (his immanence). Thus he is the All and everything and the inner Self of all beings. When God is conceived of as *brahman,* he is immutable and therefore must be the Puruṣa (cosmic Person) who is not the universe; if, however, his creation is thought to be in him, he is the world.

Accepting the *Bhāgavata-Purāṇa* as a high scriptural authority, Vaishnavism considers God the ground and subsistence of whatever exists, from whom all objects have come, by whom they continue to be, toward whom they move, and into whom they enter at the final dissolution at the end of this world, unless they already came to him in the state of emancipation (moksha). Between God and the world there is a relation of inconceivable difference in identity and identity in difference (*acintyabhedābheda;* literally, "unthinkable difference and nondifference"). The Lord creates the world merely because he wills to do so. Creation, or rather the process of differentiation and integration, is his sport (*līlā*). The world is real, but reality has two aspects: the transcendent and eternally real and the reality that is progressively realized and, in the process, bound up with the eternal aspect.

One of the chief purposes of the *Bhāgavata-Purāṇa* is the glorification of an intensely personal and passionate bhakti that gradually develops into a decidedly erotic mysticism, independent of all alternative means of salvation. According to this text, there are nine characteristics of bhakti: listening to the sin-destroying sacred histories; praising God's name; remembering and meditating on his nature and salutary endeavour (resulting in a spiritual fusion of devotee and God); serving his image; adoring him; respectful salutation; servitude; friendship; and self-surrender. Meritorious works are also an element of bhakti.

Bhagavat and Bhāgavata

According to the *Bhāgavata-Purāṇa,* the highest Bhāgavata—worshiper of the Bhagavat (God: "the Adorable One")—sees himself in all beings and all beings in the Bhagavat; free from hatred and prejudice and knowing God to be present in all beings, he loves him by loving them. Those who cannot reach this level can at least have friendly relations with coreligionists, irrespective of their birth or social status, and take compassion upon the infatuated. The true Vaishnava should worship Vishnu or one of his avatars, construct temples, bathe in holy rivers, study religious texts, serve superiors, and honour cows. In social intercourse with the adherents of other religions he tends to be passively intolerant, avoiding direct contact, without injuring them or prejudicing their rights. He should not neglect other gods but must avoid following the rituals of their followers. Misuse of the advantages of birth is severely condemned, and those who apply themselves mainly to the acquisition and enjoyment of wealth are not well qualified for bhakti. The concept of class divisions is accepted, but the idea that possession of the characteristics of a particular class is the inevitable result of birth is decidedly rejected. Because sin is antithetical to bhakti, a Brahman who is not free from falsehood, hypocrisy, envy, aggression, and pride cannot be the highest of men, and many persons of low social status may have some advantage over him in moral attitude and behaviour. The most desirable behaviour is compatible with bhakti but independent of class.

In establishing bhakti religion against any form of opposition and defending the devout irrespective of birth, the Bhāgavata religion did not actively propagate social reform; but the attempts to make religion an efficient vehicle of new spiritual and social ideas, especially Caitanya's movement, contributed, to a certain extent, to the emancipation of lowborn followers of Vishnu.

VAISHNAVISM AND ŚAIVISM

Vaishnavism. Vaishnavism is the worship of Vishnu and his various incarnations. During a long and complex development from Vedic times, there arose many Vaishnava groups with differing beliefs and aims. Some of the major Vaishnava groups include the Śrīvaiṣṇavas and Dvaitins ("philosophical or religious dualists") of South India, the followers of the teachings of Vallabha in western India, and several Vaishnava groups in Bengal in eastern India, who follow teachings derived from those of the saint Caitanya. The majority of Vaishnava believers, however, take what they like from the various traditions and blend it with various local practices.

In the Veda, Vishnu is the god of far-extending motion and pervasiveness who, for humans in distress, particularly through constrictions, penetrates and traverses the triple spaces to make their existence possible. All beings are said to dwell in his three strides or footsteps (*tri-vikrama*): his highest step, or abode, is beyond mortal ken in his dear and highest resort, the realm of heaven. So Vishnu is also the god of the pillar of the universe and is identified with the sacrifice. He imparts his all-pervading power to the sacrificer who imitates his strides and so identifies himself with the god, thus conquering the universe and attaining "the goal, the safe foundation, the highest light" (Śatapatha Brāhmaṇa).

Composite historical nature of Vishnu

In the centuries preceding the beginning of the Christian era, Vishnu became the Īśvara (immanent deity) of his special worshipers, fusing with the Puruṣa-Prajāpati figure; with Nārāyaṇa, whose cult discloses a prominent influence of ascetics; with Krishna, who in the *Bhagavadgītā* revealed a popular and universal religion, open to every-

body desiring to lead a socially normal life while having a prospect of final liberation; and with Vāsudeva, adored by a group known as the Pāñcarātras.

The extensive mythology attached to Vishnu consists largely of the mythology of his incarnations (avatars). Although the notion of "incarnation" is found elsewhere in Hinduism, it is basic to Vaishnavism. The concept is particularly geared to the social role of Vishnu; whenever dharma (universal law and order) is in danger, Vishnu departs from his heaven, Vaikuṇṭha, and incarnates himself in an earthly form to restore the good order. Each of his incarnations has a particular mythology. The classical number of these incarnations is 10, ascending from theriomorphic (animal form) to fully anthropomorphic manifestations. These are: Fish (Matsya), Tortoise (Kūrma), Boar (Varāha), Man-Lion (Narasiṃha), Dwarf (Vāmana), Rāma-with-the-Ax (Paraśurāma), King Rāma, Krishna, Buddha, and the future incarnation, Kalkin.

The 10 avatars of Vishnu

A god thus active for the good of society and the individual inspires love. Vishnu has indeed been the object of devotional religion (bhakti) to a marked degree, but mainly in his incarnations, and among them specially as Krishna and Rāma. The god rewards devotion with his grace, through which the votary may be lifted from transmigration to release. Like most other gods, Vishnu has his especial entourage: his wife is Lakṣmī or Śrī, the lotus goddess, granter of beauty, wealth, and good luck. She came forth from the ocean when gods and demons churned it in order to recover from its depths the ambrosia or elixir of immortality, *amṛta*. At the beginning of the commercial year special worship is paid to her for success in personal affairs. Vishnu's mount is the bird Garuḍa, archenemy of snakes, and his emblems are the lotus, club, discus (as a weapon), and a conch shell, which he carries in his four hands.

Whatever justification the different Vaishnava groups (*e.g.*, the Śrīvaiṣṇavas of South India or the worshipers of Vishnu Viṭhobā in Mahārāshtra) offer for their philosophical position, all Vaishnavas believe in God as a person with distinctively high qualities and worship him through his manifestations and representations. Vaishnava faith is essentially monotheistic, whether the object of adoration be Vishnu Nārāyaṇa or one of his avatars such as Rāma or Krishna. Preference for any one of these manifestations is largely a matter of tradition. Thus, most South Indian Śrīvaiṣṇavas prefer Vishnu, Rāma, or Śrī (Vishnu's consort); the North Indian groups prefer Krishna. The avatar doctrine, by accommodating the cults of various divine or heroic figures within a monotheistic framework, proved to be a powerful integrating force. Whenever the dharma declines and evil and general disaster threaten, God, the protector and preserver of the world, emanates himself and assumes an earthly form to guard the good, to destroy the wicked, and to confirm the dharma. The benevolence and beneficial activity of these figures (Rāma, Krishna, *et al.*) is, however, occasionally in doubt. In many mythical tales, Vishnu is depicted as a versatile figure of great adaptability, able, for instance, to disguise himself as a fascinating young woman in order to trick the asuras (antigods) out of the possession of the newly produced *amṛta*. His absorbing, many-sided character was a source of inspiration for various stories in which he often acts deceitfully, selfishly, or helplessly. The scene of his great deeds is usually laid in this world, especially India, in places often mentioned by name. The narratives are full of the miraculous, but their central figures give the impression of human, sometimes all too human, characters whose actions and reactions are within the limits of ordinary understanding.

The playfulness of the incarnate God

A pronounced feature of Vaishnavism is the strong tendency to devotion (bhakti), which is generally considered to be "the heart of worship," the sole true religious attitude toward a personal God, and the very foundation of the realization of man's relationship with him. Characterized by a continual consciousness of participating in God's essence, bhakti is the disinterested performance of all deeds for him, a passionate love and adoration of God, and a complete surrender to him. The widespread bhakti movement is a corollary of the Vaishnava ideal of a loving personal God and aversion to a conception of salvation that puts an end to all consciousness or individuality. Attesting to the superiority of a mystic and emotional attitude to the meditative or preponderantly ritualistic means to the highest goal, the practical and theoretical development of the bhakti idea constitutes one of the main points of difference among the several Vaishnava schools. The belief expressed in the *Bhagavadgītā*—that those who seek refuge in God with all their being will, by his benevolence and grace (*prasāda*), win peace supreme, the eternal abode—was generally accepted: bhakti will result in divine intercession with regard to the consequences of one's deeds. Among many followers of Rāmānuja, however, complete self-surrender (*prapatti*) came to be distinguished from bhakti as a superior means of spiritual realization.

Śaivism. The character and position of the Vedic god Rudra—called Śiva, "the Mild or Auspicious One," when this aspect of his ambivalent nature is emphasized—remain clearly perceptible in some of the important features of the great god Śiva, who together with Vishnu came to dominate Hinduism.

During a complex development from ancient, possibly in part from pre-Vedic, times, many different Śaiva groups arose. Major groups such as the Liṅgāyats of southern India and the Kashmir Śaivas contributed the theological principles of Śaivism, and Śaiva worship became a complex amalgam of pan-Indian Śaiva philosophy and local or folk worship.

In the minds of the ancient Indians Śiva must have been primarily the divine representative of the uncultivated, dangerous, unreliable, and much-to-be-feared aspects of nature. Śiva's character lent itself to being split into partial manifestations—each said to represent only an aspect of him—as well as to assimilating divine or demoniac powers of a similar nature from other deities. Already in the Rigveda, appeals to him for help in case of disaster—of which he might be the originator—were combined with the confirmation of his great power. In the course of the Vedic period, Śiva—originally a ritual and conceptual outsider, yet a mighty god whose benevolent aspects were readily emphasized—gradually gained access to the circle of respectable gods who preside over various spheres of human interest. Many characteristics of the Vedic Prajāpati, the creator, of Indra, the god of the phallus, and of the great Vedic god of fire, Agni, have been integrated into the figure of Śiva.

Partial manifestations of Śiva

In those circles that produced the Śvetāśvatara Upanishad (*c.* 400 BC), Śiva rose to the highest rank. Its author uses grandiose terms to show a way of escape from samsara, to proclaim Śiva the sole eternal Lord, and to establish Śiva's existence. In this description of Śiva's nature, some of the most salient features of the later Śiva, the Īśvara (immanent deity), are clearly discernible: he is the ultimate foundation of all existence and the source and ruler of all life, who, while emanating and withdrawing the universe, is the goal of that identificatory meditation that leads to complete cessation from phenomenal existence. While Vishnu became a friend nearer to man, Rudra-Śiva developed into an ambivalent and many-sided lord and master. His "doubles" or partial manifestations, however, were active among mankind: as Paśupati ("Lord of Cattle"), he took over the fetters of the Vedic Varuṇa; as Aghora ("To Whom Nothing Is Horrible"), he showed the uncanny traits of his nature (evil, death, punishment) and also their opposites.

It is not always clear in particular cases whether Śiva is invoked as a great deva (god) of frightful aspect, capable of conquering demoniac power, or as the boon-giving Lord and protector. The Īśvara idea of a Highest Being demonstrably beyond contingency is rather abstract; hence its propagators needed to use imagery, popular belief, and mythical thought. Śiva might be the sole Principle above change and variation, yet he did not sever his connections with innumerable local deities and much-feared powers worshiped by most Hindus, who still continue to invoke him in magical rites. Whereas Vishnu champions the cause of the gods, Śiva sometimes sides with the demons.

Śiva is a typical example of polarity within the Highest Being because he reconciles in his person semantically opposite though complementary aspects: he is both terri-

ble and mild, creator and agent of reabsorption, eternal rest and ceaseless activity. These seeming contradictions make him a paradoxical figure, transcending humanity and assuming a mysterious sublimity of his own. His character is so complicated and his interests are so widely divergent as to lead him in mythical narratives into conflicting situations. Yet, although Brahman philosophers like to emphasize his ascetic aspects and the ritualists of the Tantric tradition his sexuality, the seemingly opposite strands of his nature are generally accepted as two sides of one character.

Power of chastity and sexuality

Śiva interrupts his austerity and asceticism (*tapas*), which is sometimes described as continuous, to marry Pārvatī—he is even said to perform ascetic acts in order to win her love—and he combines the roles of lover and ascetic to such a degree that his wife must be an ascetic (yogi) when he devotes himself to austerities and a lustful mistress when he is in his erotic mode. This dual character finds its explanation in the ancient double conviction that unrestrained sexual intercourse is conducive to the fertility of nature and that the chastity and continence of the ascetic produce marvelous events and have an uncommon influence upon the unseen. By his very chastity, an ascetic accumulates (sexual) power that can be discharged suddenly and completely so as to produce marvelous results such as the fecundation of the soil. From various mythical tales it is seen that both chastity and the loss of chastity are necessary for fertility and the intermittent process of regeneration in nature. Ascetics engaging in erotic and creative experiences are a familiar feature in Hinduism, and the element of teeming sexuality in mythological thought counterbalances the Hindu bent for asceticism. Such sexuality, while rather idyllic in Krishna, assumes a mystical aspect in Śiva, which is why the devotee can see in him the realization of the possibilities of both asceticism and the householder state. His marriage with Pārvatī is, then, a model of conjugal love, the divine prototype of human marriage, sanctifying the forces that carry on the human race.

Śiva's myths tend to depict him as the absolutely mighty unique One, who is not responsible to anybody or for anything. His many poses express aspects of his nature: as a dancer, he is the originator of the eternal rhythm of the universe; he also catches the waters of the heavenly Ganges River, which destroy all sin; and he wears in his headdress the crescent moon, which drips the nectar of everlasting life.

Śiva represents the unpredictability of divinity. In him the Vedic Rudra is partly continued, but his mythology has become exceedingly complex. He is the hunter who slays and skins his prey and dances a wild dance while covered with the bloody hide. Far from society and the ordered world, he sits on the inaccessible Himalayan plateau of Mount Kailāsa, an austere ascetic, averse to love, who burns Kāma, the god of love, to ashes with a glance from the third eye—the eye of insight beyond duality—in the middle of his forehead. Yet another epiphany is that of the lingam, an upright rounded post, usually of stone, a formalized phallic symbol, in which form he is worshiped throughout India. And at the end of the eon, he will dance the universe to destruction. He is, nevertheless, invoked as Śiva, Śambhu, Śaṅkara (meaning: "the Auspicious One" or "the Peaceful One"), for the god that can strike down can also spare. Snakes seek his company and twine themselves around his body. He wears a necklace of skulls. He sits in meditation, with his hair braided like a hermit's, his body smeared white with ashes. These ashes recall the burning pyres on which the sannyasis (renouncers) take leave of the social order of the world and set out on a lonely course toward release, carrying with them a human skull.

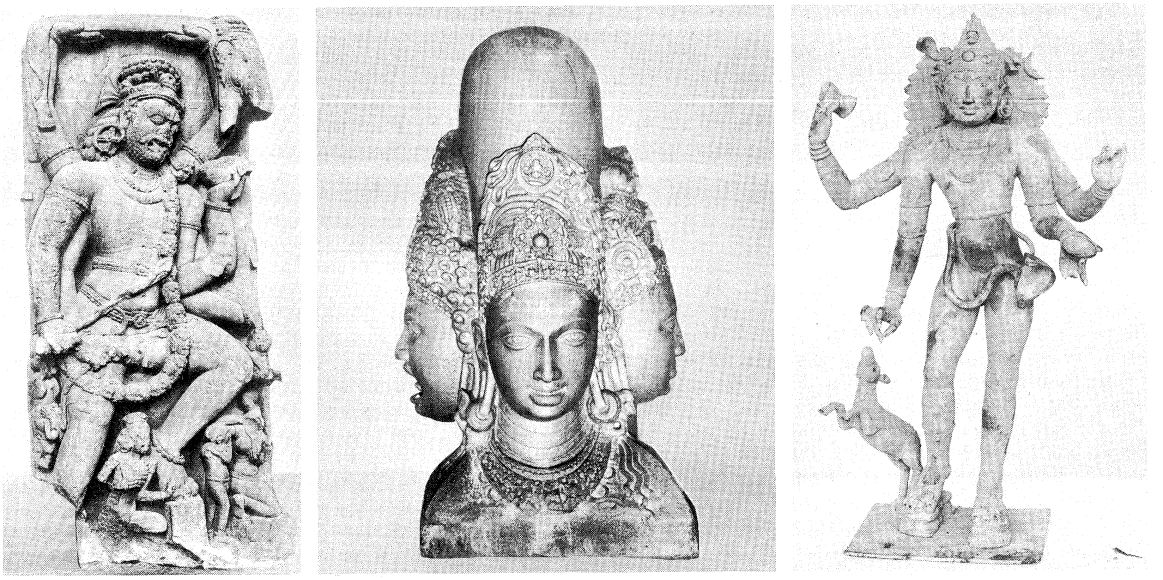
Śiva and his consorts

Like so many ascetics—often irascible and dangerous—Śiva demands to be seduced. His consort is Pārvatī ("Daughter of the Mountain"), a goddess most unlike the consorts of Vishnu in his various incarnations. She is also personified as the Goddess (Devī "goddess"), Mother (Ambā), black and destructive (Kālī), fierce (Caṇḍikā), and well-nigh inaccessible (Durgā). As Śiva's female counterpart, she inherits some of Śiva's more fearful aspects. She comes to be regarded as the power (shakti) of Śiva, without which Śiva is literally powerless. Śakti is in turn personified in the form of many different goddesses, often said to be aspects of her.

Thus the spheres of the Vishnu complex and the Śiva complex are very different ones. In important respects they represent the two different ethics of Hinduism: the dharma ethic, which aims at upholding the dharma and the cosmic and social order based on it, and the moksha (liberation) ethic, which searches for release from an order that perpetuates transmigration.

Myths of culture heroes. A culture hero can easily be assimilated to a god by identifying him with an incarnation of a god. Thus great religious teachers are considered manifestations of the god of their devotional preaching, and their lives become part of mythology. The mythology concerning great ascetics is very rich. Practically gods on Earth, these ascetics have amassed tremendous powers

Pramod Chandra



Paradoxical nature of Śiva as indicated by differing representations.
(Left) Gajāntaka, destroyer of the elephant demon, sculpture from Orissa, 8th century. In the Indian Museum, Calcutta. (Centre) Caturmukha Linga, lingam or phallic symbol, with four faces, sculpture, 5th century. In worship at Nācna-Kuṭhārā, Madhya Pradesh, India. (Right) Bhikshatana, the naked ascetic with his begging bowl, early Cōla bronze, from Tiruvengadu District, 1048. In the Thanjāvūr Art Gallery, India.

that they do not hesitate to use. The sage Kapila, meditating in the netherworld, burned to ashes 60,000 princes who had dug their way to him. Another sage, Bhagīratha, brought the Ganges River down from heaven to sanctify their ashes and, in the process, created the ocean. Agastya, revered as the Brahman who brought Sanskrit civilization to South India, drank and digested the ocean. When the Vindhya mountain range would not stop growing, Agastya crossed it to the south and commanded it to cease growing until his return; he still has not returned. Viśvāmitra, a king who became a Brahman, created a new universe with its own galaxies to spite the gods. It is in such myths that the mythopoeic imagination exults in its sensitivity to the awesome, mysterious, and marvelous.

The trial of kings

In myths concerning kings and princes, a prevailing theme is the trial of the son by the father. For example, the ancient king Yayāti had five sons to whom he wanted to transfer his own senescence for a stipulated period. All refused except the youngest, Puru. As a reward he became his father's successor, and his descendants became the Pauravas, the line of succession or dynasty in which the heroes of the *Mahābhārata* were later born. The latter heroes also underwent a trial when they were exiled from their newly won kingdom; similarly, Rāma underwent his ordeal in exile. Heroines undergo their own trials, which usually challenge their chastity, as in the case of Sītā in the *Rāmāyaṇa* and Draupadī, the one wife of all five Pāṇḍava brothers, whose sari became endless when a lustful villain attempted to pull it off.

Moving from myth to legend, there are also stories told of the great teachers, and every founder of a sect is soon deified as an incarnation of a god: the philosopher Śaṅkara (*c.* 788–820) as an incarnation of Śiva, the religious leader Rāmānuja (d. AD 1137) as that of Nārāyaṇa-Vishnu, and the Bengal teacher Caitanya (1485–1533) simultaneously as that of Krishna and his beloved Rādhā.

Myths of holy rivers and places. Of particular sanctity in India are the perennial rivers, among which the Ganges stands first. This river, personified as a goddess, originally flowed only in heaven until she was brought down by Bhagīratha to purify the ashes of his ancestors. She came down reluctantly, cascading first on the head of Śiva, in order to break her fall, which would have shattered the Earth. Confluences are particularly holy, and the Ganges' confluence with the Yamuna at Allahābād is the most sacred spot in India. Another river of importance is the Sarasvatī, which loses itself in desert; it was personified as a goddess of eloquence and learning.

Every major and many minor temples and sanctuaries have their own myths of how they were founded and what miracles were wrought there. The same is true of famous places of pilgrimage, usually at sacred spots near and in rivers; important among these are Vrindāvana (Brindāban) on the Yamuna, which is held to be the scene of the youthful adventures of Krishna and the cowherd wives. Another such centre with its own myths is Gayā, especially sacred for the funerary rites that are held there. And there is no spot in Vārānasi (Benares) along the Ganges that is without its own mythical history.

PHILOSOPHICAL TEXTS

Although the details of Indian philosophy, as it was developed by professional philosophers, may be treated as a subject separate from Hinduism (see INDIAN PHILOSOPHY), certain broad philosophical concepts were absorbed into the myths and rituals of Hindus and are best viewed as a component of the religious tradition.

Mysticism. One of the major trends of Indian religious philosophy is a kind of mysticism: the desire for union of the self with something greater than the self, whether that be defined as a principle that pervades the universe or as a personal God. Hindu mysticism includes both these forms and a great many that lie in between. At one extreme is the realization of the identity of the individual self with the impersonal principle called *brahman,* the position of the Vedānta school of Indian philosophy; and at the other is the intensive devotionalism to a personal God, called by a variety of names, that is found in the bhakti (devotional) sects.

There are four things common to most Hindu mystical thought. First, it is based on experience: the state of realization, whatever it is called, is both knowable and communicable, and the systems are all designed to teach people how to reach it. It is not, in other words, pure speculation. Second, it has as its goal the release of the spirit-substance of the individual from its prison in matter, whether matter be considered real or illusory. Matter is the cause of the suffering of which Buddhism speaks. Third, all the systems recognize the importance or the necessity of the control of the mind and body as a means of realization; sometimes this takes the form of extreme asceticism and mortification, and sometimes, at the other extreme, it takes the form of the cultivation of mind and body in order that their energies may be properly channeled. And, finally, at the core of Hindu mystical thought is the functional principle that knowing is being. Thus, knowledge is something more than analytical categorizing: it is total understanding. This understanding can be purely intellectual, and some schools equate the final goal with omniscience, as does yoga. Knowing can also mean total transformation: if one truly knows something, he is that thing. Thus, in the devotional schools, the goal of the devotee is to transform himself into a being who, in eternity, is in immediate and loving relationship to the deity. But despite the fact that these are both ways of knowing, the difference between them is significant. In the first instance, the individual has the responsibility to train and use his own intellect. The love relationship of the second, on the other hand, is one of dependence, and the deity assists the devotee through grace. The distinction is generally made by the analogy of the cat and the monkey: the cat carries her young in her mouth, and thus the kitten has no responsibility. But the young monkey must cling by its own strength to its mother's back.

Knowledge as being and becoming

It is usual for writers on the subject, following Surendranath Dasgupta, a historian of Indian philosophy, to list five major varieties of Hindu mysticism, the five having arisen in historical order as follows:

1. The sacrificial, based on the Vedas and Brāhmaṇas.
2. The Upanishadic, in which are found the beginnings of both monistic (concerned with a unitary principle of reality, immanent in the world) and theistic (concerned with a personal or suprapersonal God) systems.
3. The yogic, relating to physical and mental discipline; the earliest known text of this school is the *Yoga-sūtra* of Patañjali, dated variously between the 2nd century BC and the 5th century AD. According to yogic mysticism, man realizes union by means of physical and mental control of himself, which in turn leads to control of both natural and divine forces.
4. The Buddhistic, in which enlightenment is the realization of the four Truths—the fact of suffering, the cause of suffering, the cessation of suffering, and the means of arriving at these three truths: the Eightfold Path. The ultimate state, the culmination of one path of the Eightfold Path, is nirvana, "the blowing out," the extinction of desire (see BUDDHISM, THE BUDDHA AND).
5. The devotional, or bhakti, type of mysticism comprises a range of theistic systems, with a conception of absolute dualism between man and God on the one extreme, and a conception of qualified nondualism on the other. Although there are traces of this devotionalism throughout the history of Indian religion, it began to develop in earnest in South India in the 7th through 10th century AD with the hymns of the poet saints called Āḻvārs.

Philosophical sutras and the rise of the six schools of philosophy. From about the beginning of the Christian era through the period of the Gupta empire, the systems of the Six Schools (Ṣaḍdarśana) of orthodox philosophy were formulated in terse sutras.

Śaṅkara and the Vedānta school

The most important of the Six Schools is the Vedānta ("End of the Vedas"), also called Uttara-Mīmāṃsa, or later Mīmāṃsā. The most renowned philosopher of this school and, indeed, of all Hinduism was Śaṅkara (traditionally dated *c.* 788–820, but he probably died about 20 years later). He was born at Kālaḍi in Kerala and is said to have spent most of his life traveling through India debating with members of other sects. The Śaṅkaran system

has sounded the keynote of intellectual Hinduism down to the present, but later teachers founded sub-schools of Vedānta, which are perhaps equally important.

Śaṅkara was also responsible for the growth of Hindu monasticism, which had been in existence for more than a millennium in the form of hermit colonies. Inscriptions from Gupta times onward also refer to monklike orders of Śaiva ascetics, apparently living according to distinctive disciplines and with distinguishing garments and emblems. Śaṅkara founded a closely disciplined Śaiva order, perhaps partly modeled on the Buddhist sangha, or order, known as *daśnāmī,* which is still the most influential orthodox Hindu ascetic group. The order is composed of 10 brotherhoods and hence called *daśnāmīs* ("those with 10 names"). Orders became an established institution with wider geographic affiliations. Some of these admitted Brahmans only; others were open to all four classes or even to women; some made a practice of nudity. These Śaiva communities are more inclined to individual asceticism and are less closely organized than the Vaishnava *vairāgins* ("the dispassionate") or *gosvāmins* ("the masters"). Śaṅkara is also said to have founded the four main monasteries (*maṭha*) at the four corners of India: Sringeri in Karnataka, Badrīnāth in the Himalayas, Dwārkā in Gujarāt, and Puri in Orissa. The abbots of these monasteries control the spiritual lives of many millions of devout Śaiva laymen throughout India, and their establishments strive to maintain the traditional philosophical Hinduism of the strict Vedānta. In modern times, certain *daśnāmī* leaders have incurred criticism for their firm opposition to social change.

Vaishnava theology of Rāmānuja

The theologians had to assume the task of explaining the relation between God, as the unaffected and unchanging cause of all things, and the universe. According to Rāmānuja (*c.* 1017–1137), a great South Indian thinker of Śrīvaiṣṇava persuasion, *brahman* (*i.e.,* God) is a Person with high attributes, the object of an individual's search for the higher knowledge that is the only entrance to salvation. Because an absolute creation is denied, God is viewed as the sole cause of his own modifications; namely, the emanation, existence, and absorption of the universe. Although unlimitedly expansive, God is conceived to be essentially different from everything material, the absolute opposite of any evil, free from any imperfection, omniscient, omnipotent, possessed of all positive qualities (such as knowledge, bliss, beauty, and truth), of incomparable majesty, the inner soul of all beings, and the ultimate goal of every religious effort. The universe is considered a real transformation of *brahman,* whose "body" consists of the conscious souls and everything unconscious in their subtle and gross states. The karma doctrine is modified as follows: the Lord, having determined good and bad deeds, provides all individual souls with a body in which they perform deeds, reveals to them the scriptures from which they may learn the dharma, and enters into them as their internal regulator. The individual acts at his own discretion but needs the Lord's assent. If the devotee wishes to please him, God induces him, with infallible justice and loving regard, to intentions and effort to perform good deeds by which the devotee will attain to him; if not, God keeps him from that goal.

The influence of the bhakti movement had earlier led Rāmānuja to admit a twofold possibility of emancipation: in addition to the meditative method of the highest insight (jnana) into the oneness of soul and God, which destroys the residues of karma and propitiates God to win his grace, there is the way of bhakti. Those who prefer the former way will reach a state of isolation, the others an infinitely blissful eternal life in, through, and for God, with whom they are one in nature but not identical. They do not lose their individuality and may even meet Vishnu in his Vaikuṇṭha heaven and enjoy delight beyond description.

An interesting development of Rāmānuja's doctrine of qualified monism is found in the philosophy of Madhva (died *c.* 1276). This Kanarese Brahman taught a doctrine of dualism according to which God and the soul are eternally distinct.

Authors of Śaiva *Purāṇas* established two ingenious and complementary doctrines to explain the nature and omnipotence of God (the force that rules, absorbs, and reproduces the world and that in performing any one of these acts necessarily performs the other two as well), the existence of the world, and the identity of God and the world. A theory of five "faces," or manifestations—each of which is given mythological names and related mantras—is of great ritual significance. It associates Śiva's so-called creative function, by which he provokes the evolution of the material cause of the universe, with his first face, or aspect; its maintenance and reabsorption with his second and third faces; his power of obscuration, by which he conceals the souls in the phenomena of samsara, with his fourth face; and his ability to bestow his grace, which leads to final emancipation, with his fifth face. The five functions are an emanation of the unmanifested Śiva who is the transcendent *brahman.*

The faces became the central elements of a comprehensive classification system. They were identified with parts of God's body, regions of the universe, various ontological principles, organs of sense and action, and the elements. The system was used to explain how Śiva's being is the All and how the universe is exclusively composed of aspects or manifestations of Śiva. In his fivefold nature, Śiva was shown to be identical with the 25 (five times five) elements or principles assumed by the prominent Saṃkhyā school of Indian philosophy. The special significance of the number five in Śaivism can be understood as a philosophical elaboration of the time-honoured fourfold organization of the universe. (The four quarters of the sky also play a prominent part in religious practice.) According to this conception, a fifth aspect, when added to the four, is considered the most important aspect of the group because it represents each of the four and collectively unites all of their functions in itself. The system finds its complement in the doctrine of the five Sadākhyas (five items that bear the name sat, "is" or "being") representing the five aspects of that state, which may be spoken of as the experience of "there is" (sat) and which have evolved from God's fivefold creative energy (shakti). In these, God "dwells" in his aspect called Sadāśiva ("the Eternal Śiva"), which is regarded sometimes as a manifestation of and sometimes as identical with the Supreme Being.

Another Śaiva doctrine posits eight "embodiments" of Śiva as the elements of nature (ether, wind, fire, water, earth), Sun, Moon, and the sacrificer, or consecrated worshiper (also called Atman). To each of these eight elements corresponds one of Śiva's traditional names or aspects—to the last one, usually Paśupati. The world is a product of these eight forms, consists of them, and can only exist and fulfill its task because the eight embodiments cooperate. Because each individual is also composed of the same eight realities (*e.g.,* the light of man's eyes corresponds to that of the Sun), Śiva constitutes the corporeal frame and the psychical organism of every living being. The eighth constituent is the indispensable performer of the rites that sustain the gods who preside over the cosmic processes and are really Śiva's faculties.

Regional variations in Śaivism

Although Śaivism is a much more coherent whole than Vaishnavism, there evolved, in different parts of India, some branches with peculiarities of their own. According to the pronouncedly idealist monism of Kashmir Śaivism, an important religiophilosophic school, Śiva manifests himself through a special power as the first cause of creation, and he also manifests himself through a second power as the innumerable individual souls who, because of a veil of impurity, forget that they are the embodiment of the Highest. This veil can be torn off by intense faith and constant meditation on God, by which the soul transmutes itself into a universal soul and eventually attains liberation through a lightninglike, intuitive insight into its own nature. Those Hindus who adhere to this group consider their doctrine a manifestation of the highest Reality, Knowing Consciousness, neither personal nor impersonal; as Śiva in the form of the transcendent Word, which is his unspoken Thought, the content of which is the universe.

The Śaiva-siddhānta, a prominent religiophilosophic school of Tamil-speaking South India, assumes three eternal principles: God (who is independent existence, unqualified intelligence, and absolute bliss), the universe, and the

souls. The world, because it is created by God (efficient cause) through his conscious power (instrumental cause) and maya (material cause), is no illusion. The main purpose of its creation is the liberation of the beginningless souls, which are conceived as "cattle" (*paśu*) bound by the noose (*pāśa*) of impurity (*mala*) or spiritual ignorance, which forces them to produce karma. However, they see the karma process as a benefit, for as soon as the soul has sufficiently ripened and reached a state of purity enabling it to strive after the highest insight, God graciously intervenes, appearing in the shape of a fully qualified and liberated spiritual guide (guru), through whose words God permits himself to be realized by the individual soul.

TANTRISM

Tantric traditions and Śāktism. Toward the end of the 5th century, the cult of the Mother Goddess began to achieve a significant place in religious life. Śāktism, the worship of the Śakti, the active power of the godhead conceived in feminine terms, should be distinguished from Tantrism, the search for spiritual power and ultimate release by means of the repetition of sacred syllables and phrases (mantras), symbolic drawings (mandalas), and other secret rites elaborated in the texts known as tantras ("looms").

In many respects the tantras are similar to the *Purāṇas*. Theoretically a tantra deals with (1) knowledge, or philosophy, (2) yoga, or concentration techniques, (3) ritual, which includes the formation of icons and the building of temples, and (4) conduct in religious worship and social practice. In general the last two subjects preponderate, while yoga tends to centre on the mystique of certain sound-symbols (mantras) that sum up esoteric doctrines. The philosophy tends to be a syncretistic mixture of Sāṃkhya and Vedānta philosophical thought, with special and at times exclusive emphasis on the god's power, or shakti. The Tantric texts can be divided into three classes: (1) Śaiva *Āgamas* (traditions of the followers of Śiva), (2) Vaishnava *Saṃhitās* ("Collections of the Vaishnavas," a name borrowed from the Vedic Saṃhitās), and (3) Śākta Tantras ("Looms of the Followers of the Goddess Śakti"). However, they all have the common bond of venerating the Goddess.

Śaiva, Vaishnava, and Śākta tantras

Surviving Hindu tantras were written much later than many of those of Tantric Buddhism, and it may be that the Hindus derived much from the Buddhists in this respect. Although there is early evidence of Tantrism and Śāktism in other parts of India, the chief centres of both were modern Bengal, Bihār, and Assam.

Śaiva Āgamas. Like much other Hindu sacred literature, this literature is neither well-cataloged nor thoroughly studied. It is only possible here to summarize classes of texts within the various traditions.

The sects of Āgamic Śaivas (Śiva worshipers who follow their own Āgama—"traditional"—texts) encompass both the Sanskritic Śaiva-siddhānta—*i.e.,* those who accept the philosophical premises and conclusions of Śaivas in the north—and the southern Liṅgāyats or Vīraśaivas (from *vīra,* literally "hero"; a lingam is the Śiva emblem that is worshiped in lieu of images). The Śaiva-siddhānta traditionally has 28 *Āgamas* and 150 sub-*Āgamas*. Their principal texts are hard to date; most probably they do not antedate the 8th century. Their doctrine states that Śiva is the conscious principle of the universe, while matter is unconscious. Śiva's power, or shakti, personified as a goddess, causes bondage and release. She is also the magic Word, and thus her nature can be sought out and meditated upon in mantras.

Kashmir Śaivism begins with the *Śivasūtra* or "Lines of Doctrine Concerning Śiva" (*c.* 850) as a new revelation of Śiva. The system embraces the *Śivadṛṣṭi* ("A Vision of Śiva") of Somānanda (950), in which emphasis is placed on the continuous recognition of Śiva; the world is a manifestation of Śiva brought about by his shakti. The system is called *trika* ("triad"), because it recognizes the three principles of Śiva, Śakti, and the individual soul. Vīraśaiva texts begin at about 1150 with the *Vācanams* ("Sayings") of Basava. The sect is puritanical, worships Śiva exclusively, rejects the caste system in favour of its own social organization, and is highly structured with monasteries and gurus.

Vaishnava Saṃhitās. These consist of two groups of texts: *Vaikhānasa Saṃhitās* and *Pāñcarātra Saṃhitās*. The latter group is the prevailing one; more than 200 titles are known, though the official number is 108. *Vaikhānasa Saṃhitās* (collections of the Vaishnava school of *Vaikhānasas*, who were originally ascetics) seem to have embodied the original temple manuals for the Bhāgavatas (devotees), which by the 11th or 12th century had become supplanted by the *Pāñcarātra Saṃhitās* (collections of the Vaishnava school of Pāñcarātra—"the System of the Five Nights"). The philosophy of the latter is largely a matter of cosmogony, greatly inspired by both the Sāṃkhya and yoga philosophies.

Notion of shakti in Vaishnavism

Apart from their theology, in which for the first time the notion of shakti is introduced into Vaishnavism, they are important because they give an exposition of Vaishnava temple and cult practices. On the philosophical side it is maintained that the supreme god Krishna Vāsudeva manifests himself in four coequal "divisions" (*vyūhas*), representing levels in creation. These gods emanate as supramundane patrons before the primary creation is started by their shakti (power). In the primary creation Śakti manifests herself as a female creative force inspired by the Sāṃkhya philosophy's cosmogony. Practically, stress is laid on a type of incarnation—"iconic incarnation"—in which the god is actually present with a portion of himself in a stone or statue, which thus becomes an icon; therefore the icon can be worshiped as God himself.

Śākta Tantras. Śāktism in one form or the other has been known since Bāṇa (*c.* 650) wrote his *Hundred Couplets to Caṇḍi* (*Caṇḍi-śataka*) and Bhavabhūti his play *Mālatī Mādhava* (725), both of which refer to Tantric practices. There is no traditional authoritative list of texts, but many texts are extant.

Śāktism is an amalgam of Śaivism and folk mother-goddess cults. The Śaiva notion that not Śiva himself but his shakti (sexual, creative power) is active is taken to the extreme—that, without Śakti, Śiva is a corpse, and simultaneously that Śakti is the creator as well as creation. In yoga, great importance is ascribed to mantras, which conjure up the realities with which they are identified. Another important notion (partly derived from yoga philosophy) is that through the body run subtle canals that carry esoteric powers connected with the spinal cord, at the bottom of which the Goddess is coiled around the lingam as Kuṇḍalinī; she can be made to rise through the body to the top, whereupon release from samsara takes place. Important among the Śākta Tantras are the *Kulārṇava Tantra* ("Ocean of Tantrism"), which gives details on the "left-handed" cult forms of ritual copulation; the *Kulacūḍāmaṇi* ("Crown Jewel of Tantrism"), which embroiders on ritual; and the *Śaradātilaka* ("Beauty Mark of the Goddess Śaradā") of Lakṣmaṇadeśika (11th century), which discusses magic.

A temple was erected in honour of the mother goddesses at Gangdhār, Rājasthān, in AD 423. There, magical rites of a terrifying kind were practiced, for the temple is described as "loud with the shouts of demonesses, crying in the thick darkness," by the playwright Bhavabhūti, whose drama *Mālatī Mādhava* (about the adventures of the hero Mādhava and his beloved Mālatī) contains a scene depicting secret rites with human sacrifice and ritual cannibalism. The goddess cults eventually centred around Durgā, the consort of Śiva, in her fiercer aspect.

Nature of Tantric tradition. Tantrism, which appears both in Buddhism and in Hinduism, is an important component of religion that, though primarily meant for esoteric circles, also influenced, from the 5th century AD, many religious trends and movements. Opinions of what Tantrism is are quite diverse. Generally, Tantrism claims to show in times of religious decadence a new way to the highest goal and bases itself upon mystic speculations concerning divine creative energy (shakti). Tantrism is a method of conquering transcendent powers and realizing oneness with the highest principle by yogic and ritual means—in part magical and orgiastic—which are also supposed to achieve other supranormal goals.

Tantrists take for granted that all factors in both the macrocosm and the microcosm are closely connected. The adept (*sādhaka*) has to perform the relevant rites on his own body, transforming its normal, chaotic state into a "cosmos." The macrocosm is conceived as a complex system of powers that by means of ritual-psychological techniques can be activated and organized within the individual body of the adept. Contrary to the ascetic emancipation methods of other groups, the Tantrists emphasize the activation and sublimation of the possibilities of their own body, without which salvation is believed to be beyond reach.

Vāmācāra: left-handed Tantra

The Tantrists of the *Vāmācāra* ("the left-hand practice") sought to intensify their own sense impressions by making enjoyment, or sensuality (*bhoga*), their principal concern: the adept pursued his spiritual objective through his natural functions and inclinations, which were sublimated and then gratified in rituals in order to disintegrate his normal personality. This implies that cultic life was also largely interiorized and that the whole world, because it became completely ritualized, was given a new and esoteric meaning.

Tantric worship (puja) is complicated and in many respects different from the conventional ceremonies that it has often influenced. Tantric devotees distinguish between an "external" and an esoteric meaning of their texts and interpret their texts by means of an ambiguous "twilight" language. Tantrists describe states of consciousness with erotic terminology and describe physiological processes with cosmological terminology. They proceed from "external" to "internal" worship and adore the Goddess mentally, offering their hearts as her throne and their self-renunciation as "flowers."

According to Tantrism, concentration is intended to evoke an internal image of the deity and to resuscitate the powers inherent in it so that the symbol changes into mental experience. This "symbolic ambiguity" is also much in evidence in the esoteric interpretation of ritual acts performed in connection with images, flowers, and other cult objects and is intended to bring about a transfiguration in the mind of the adept.

The mantras (sacred utterances, such as *hūṃ, hrīṃ,* and *klaṃ*) are an indispensable means of entering into contact with the power they bear and of transcending normal mundane existence. Most potent are the monosyllabic, fundamental, so-called *bīja* ("seed") mantras, which constitute the main element of longer formulas and embody the essence of divine power as the eternal, indestructible prototypes from which anything phenomenal derives its existence. The cosmos itself owes its very structure and harmony to them. Also important is the introduction of spiritual qualities or divine power into the body (*nyāsa*) by placing a finger on the relevant spot (accompanied by a mantra).

Those Tantrists who follow the "right-hand path" attach much value to the yoga that developed under their influence and to bhakti and aspire to union with the Supreme by emotional-dynamic means, their yoga being a self-abnegation in order to reach a state of ecstatic blissfulness in which the passive soul is lifted up by divine grace.

There is also a Tantric mantra yoga (discipline through spells), which operates with formulas, and a hatha-yoga, (Sanskrit: "union of force"). In addition to normal yogic practices—abstinences, observances, bodily postures, breath control that requires intensive training, withdrawal of the mind from external objects, concentration, contemplation, and identification that are technically helped by mudras (*i.e.,* ritual intertwining of fingers or gestures expressing the metaphysical aspect of the ceremonies or the transformation effected by the mantras) and muscular contractions—hatha-yoga consists of internal purifications (*e.g.,* washing out stomach and bowels), shaking the abdomen, and some forms of self-torture. The whole process is intended to "control the 'gross body' in order to free the 'subtle body.' "

Mudras

Some Tantrists also employ *laya yoga* ("reintegration by mergence"), in which the female nature-energy (representing the shakti), which is said to remain dormant and coiled in the form of a serpent (*kuṇḍalinī*) representing the uncreated, is awakened and made to rise through the six centres (chakras) of the body, which are located along the central artery of the subtle body, from the root centre to the lotus of a thousand petals at the top of the head, where it merges into the Puruṣa, the male Supreme Being. As soon as the union of shakti and Puruṣa has become permanent, according to this doctrine, wonderful visions and powers come to the adept, who then is emancipated. Some of the Tantric texts also pursue worldly objectives involving magic or medicine.

Tantric and Śākta views of nature, man, and the sacred. The Tantric movement is sometimes inextricably interwoven with Śāktism. Śāktism consists of doctrines and practices that assume the existence of one or more shaktis. These are "creative energies" that are inherent in and proceed from God and are also capable of being imagined as female deities. Shakti is the deciding factor in

Shakti

By courtesy of (centre) the Cleveland Museum of Art, Ohio, Mr. and Mrs. William H. Marlatt Fund; photographs, (left and right) Pramod Chandra



Various roles of Śakti, the female aspect of the divine.
(Left) Pārvatī, the beneficent mother, bronze statue, south Indian, *c.* 900. Formerly in the collection of Srinivasa Gopalachari. (Centre) Durgā, the destroyer, detail from a Basohli school painting, *c.* 1700. In the Cleveland Museum of Art, Ohio. (Right) Ardhanāriśvara, united with lord Śiva as half-male, half-female, sandstone sculpture from Jhālawār, 6th century. In the Jhālawār Archaeology Museum, Rājasthān, India.

the salvation of the individual and in the processes of the universe because God acts only through his energy, which, personified as a goddess, is his spouse. Her role is very different in the various systems: she may be considered the central figure in a philosophically established doctrine, the dynamic aspect of *brahman,* producing the universe through her maya, or mysterious power of illusion; a capricious demoniac ruler of nature in its destructive aspects; a benign mother goddess; or the queen of a celestial court. There is a comprehensive Śāktism that identifies the goddess (usually Durgā) with *brahman* and worships her as the ruler of the universe by virtue of whom even Śiva exists. As Mahāyoginī ("Great Mistress of Yoga"), she produces, maintains, and reabsorbs the world. As the eternal Mother, she is exalted in the *Devīmāhātmya* ("Glorification of the Goddess") section of the *Mārkaṇḍeya-Purāṇa* (an important medieval Śākta encyclopaedic text). In the Bengal cult of the goddess Kālī, she demands bloody sacrifices from her worshipers lest her creative potency fail her; this cult propounds the belief that birth and death are inseparable, that joy and grief spring from the same source, and that the frightening manifestations of the divine should be faced calmly.

The Vedic goddess Vāc (her name means "Word") was then already the energetic and productive partner of Prajāpati. As Ardhanārīśvara (the "Lord Who Is Half Female"), Śiva presides over procreation. The Śāktas—often markedly associated with Śaivism—drew the following conclusions: creation is the result of the eternal lust of the divine couple; the man who is blissfully embraced by a beloved woman who is Pārvatī's counterpart assumes Śiva's wonderful personality and, liberated, will continue the joy of amorous sport.

Vaishnava Śāktism

In all of his incarnations Vishnu is united with his consort, Lakṣmī. The sacred tales of his various relations with her manifestations led his worshipers to view human devotion as parallel to the divine love and hence as universal, eternal, and sanctified. In Vaishnava Tantrism, Lakṣmī plays an important part as God's shakti—that is, as a central metaphysical principle. In his supreme state, Vishnu and his shakti are indissolubly associated with one another so as to constitute the personal manifestation of the supreme *brahman,* also called Lakṣmī-Nārāyaṇa. In mythical imagery, Lakṣmī never leaves Vishnu's bosom. In the first stage of creation, she awakens in her dual aspect of action-and-becoming, in which she is the instrumental and material cause of the universe; Vishnu himself is the efficient cause. In the second stage, her "becoming" aspect is manifested in the grosser forms of the souls and the power of maya, which is the immaterial source of the universe. In displaying her power she takes into consideration the accumulated karma of the beings, judging mundane existence as merit and demerit. Presented in myth as God's wife and the queen of the universe, she is always intent on liberating, by her favour and compassion, the incarnated souls of the devout; that is, she allows them to reenter into herself because they are really "parts," or rather "contractions," of her own essence. After entering her, the liberated soul takes part in the perfect embrace of the divine couple. Pāñcarātra Vaishnavism emphasizes that Lakṣmī—who in the mythological sphere intercedes with her husband for the preservation of the world—spontaneously and by virtue of her own power differentiates herself from Vishnu because she has in view the liberation of the souls. This current of thought complicated its explanation of the relation between God and the universe—which was at the same time an attempt at assigning to God's manifestations a place in a harmonious theological and cosmological system—with an evolutionist theory of successive creations. God is assumed to manifest himself also in three other figures, mythologically his brothers, who, each with his own responsibility, have not only a creative but also an ethical function, by which they assist those who seek to attain to final emancipation.

Tantric ritual and magical practices. The ritual of the left-hand Tantrists consisted of a kind of black mass in which all of the taboos of conventional Hinduism were conscientiously violated. Thus, in place of the traditional five elements (*tattvas*) of the Hindu cosmos, these Tantrists used the five "m"s: *māṃsa* (flesh, meat), *matsya* (fish), *madya* (fermented grapes, wine), *mudrā* (frumentum, cereal, parched grain, or gestures), and *maithuna* (fornication). This latter element was made particularly antinomian through the involvement of forbidden women, such as one's sister, mother, the wife of another man, or a low-caste woman, who was identified with the Goddess. Menstrual blood, strictly taboo in conventional Hinduism, was also used at times. Such rituals, which are described in Tantric texts and in tracts against Tantrists, made the Tantrists notorious. It is likely, however, that the rituals were not regularly performed except by a relatively small group of highly trained adepts; the usual Tantric ceremony was purely symbolic and even more fastidious than the pujas in Hindu temples.

The cult of the Śāktas is based on the principle of the ritual sublimation of natural impulses to maintain and reproduce life. Śākta adepts are trained to direct all their energies toward the conquest of the Eternal. The ritual satisfaction of lust and the consumption of consecrated meat or liquor are esoterically significant means of realizing the unity of flesh and spirit, of the human and the divine. They are not considered sinful acts but, on the contrary, effective means of salvation. Ritual copulation—which may also be accomplished symbolically—is, for both partners, a form of sacralization, the act being a participation in cosmic and divine processes. The experience of transcending space and time, of surpassing the phenomenal duality of spirit and matter, of recovering the primeval unity, the realization of the identity of God and his Śakti, and of the manifested and unmanifested aspects of the All, constitute the very mystery of Śāktism. The interpretation—metaphorical or literal—of the doctrines is, however, largely a matter of opinion and practice. Ritual practice is indeed as varied as the doctrines. Extreme Śākta communities perform the secret nocturnal rites of the *śrīcakra* ("wheel of radiance," described in the *Kulārṇava Tantra*), in which they avail themselves of the natural and esoteric symbolic properties of colours, sounds, and perfumes to intensify their sexual experiences. Most Tantrists, however, eliminate all but the verbal ritual.

Individual and collective yoga and worship, conducted daily, fortnightly, and monthly "for the delectation of the deity," are of special importance. After elaborate purifications, the worshipers—who must be initiated, full of devotion toward the guru and God, have control over themselves, be well prepared and pure of heart, know the mysteries of the scriptures, and look forward to the adoration with eagerness—make the prescribed offerings, worship the mighty puissance of the Divine Mother, and recite the relevant mantras. Once they have become aware of their own state of divinity, they are qualified to unite sexually with the Goddess. If a woman is, in certain rituals, made the object of sexual worship, the Goddess is first invoked into her; the worshiper is not to cohabit with her until his mind is free from impurity and he has risen to divine status. Copulation with a low-caste woman helps to transcend all opposites; union with a woman who belongs to another man is often preferred because it is harder to obtain, nothing is certain in it, and the longing stemming from the separation of lover and beloved is more intense—it is pure *preman* (agape, or divine love); adoration of a girl of 16 aims at securing the completeness and perfection of which this number is said to be the expression. However, the texts reiterate how dangerous these rites are for those who are not initiated; those who perform such ritual acts without merging their minds in the Supreme are likely to go to one of the hells.

Bengali Vaishnava-Sahajiyā cult

The esoteric Vaishnava-Sahajiyā cult, which arose in Bengal in the 16th century, was another emotional attempt at reconciling the spirit and the flesh. Displaying contempt for social opinion, its adherents, using the natural (*sahaja*) qualities of the senses and stressing the sexual symbolism of Bengal Vaishnavism, reinterpreted the Rādhā–Krishna legend and sought for the perpetual experience of divine joy: because Krishna's nature is love and the giving of love and because man is identical with Krishna, the realization of love can, after an arduous training, be experienced in man's nature. Women, being a ritual necessity as well

as the embodiment of a theological principle, could even become spiritual guides, like Rādhā, conducting the worshiper in his search for realization. After reaching this state, he remains in eternal bliss, can dispense with guru and ritual, and be completely indifferent to the world, "steadfast amidst the dance of maya."

Tantric and Śākta ethical and social doctrines. These ethical and social principles, though fundamentally the same as those promulgated in the classical dharma works, breathe a spirit of liberality: much value is set upon family life and respect for women (the image of the Goddess); no ban is placed on traveling (conventionally regarded as bringing about ritual pollution) or on the remarriage of widows. Although Tantric and Śākta traditions did not oblige their followers to deviate in a socially visible way from the established order, they provided a ritual and a way of life for those who, because of sex or caste, could not participate satisfyingly in the conventional rites.

The ancient Tantric tradition, based on the esoteric tantra literature, has become, through time, so interwoven with more orthodox Hinduism that it is difficult to define precisely. Although it sees an identity between the soul and the cosmos, it speaks of the internalization of the cosmos rather than of the release of the soul to its natural state of unity. The body is the microcosm, and the ultimate state is not only omniscience but total realization of all universal and eternal forces. The body is real, not because it is the function or creation of a real deity but because it contains the deity, together with the rest of the universe. The individual soul does not unite with the One—it is the One, and the body is its function.

Tantrism, though not always in its full esoteric form, is a feature of much modern mystical thought. In Tantrism the consciousness is spoken of as moving—driven by repetition of the mantra and by other disciplines—from gross awareness of the material world to realization of the ultimate unity. The image is of a serpent, coiled and dormant, awakened and driven upward in the body through various stages of enlightenment until it reaches the brain, the highest awareness. The modern mystic Ramakrishna describes the process, which also describes the experience that all Hindu mystical processes seek:

The coiled serpent

> When [the serpent] is awakened, it passes gradually through [various stages], and comes to rest in the heart. Then the mind moves away from [the gross physical senses]; there is perception, and a great brilliance is seen. The worshiper, when he sees this brilliance, is struck with wonder. The [serpent] moves thus through six stages, and coming to [the highest one], is united with it. Then there is *samādhi*... When [the serpent] rises to the sixth stage, the form of God is seen. But a slight veil remains; it is as if one sees a light within a lantern, and thinks that the light itself can be touched, but the glass intervenes.... In *samādhi,* nothing external remains. One cannot even take care of his body any more; if milk is put into his mouth, he cannot swallow. If he remains for twenty-one days in this condition, he is dead. The ship puts out to sea, and returns no more. (Translation by Edward C. Dimock, Jr. Source is *Śrīśrīrāmakṛṣṇa-kathāmṛta;* Calcutta: Ramakrsna Mission.)

VERNACULAR LITERATURES

Most of the texts cited in this survey are Sanskrit texts, which constitute the oldest layer of preserved Hindu literature. But the sacred literature of India is by no means as monolithic as these texts might suggest. Several other essential elements exist: independent sacred literatures in languages other than Sanskrit and material in other languages related to the Sanskrit texts either as sources of material now preserved only in Sanskrit or as new texts originating as translations of Sanskrit texts. Because Sanskrit has been in intimate contact with the "mother tongues" of India for such a long time, it is often impossible to determine in which of these categories a particular vernacular text belongs.

Indologists usually emphasize the influence of Sanskritic (often called "Aryan") culture on Dravidian culture, and indeed this influence was considerable. Sanskritic influence was already in evidence in the earliest Tamil (a principal Dravidian language) literature, perhaps dating from the 1st century AD. At this time in South India the orthodox cults were aristocratic in character and were supported by kings and chiefs who gained in prestige by patronizing Brahmans and adopting Aryan ways. The Tamils were still primarily devoted to the old cults, some of which, however, were taking on an Aryan complexion. The pastoral god Murugaṉ was identified with Skanda and his mother, the fierce war goddess Koṟṟavai, with Durgā. Varuṇaṉ, a sea god who had adopted the name of the old Vedic god but otherwise had few Aryan features, and Māyoṉ, a black god who was a rural divinity with many of the characteristics of Krishna in his pastoral aspect, also are depicted in Tamil literature. The final Sanskritization of the Tamils was brought about through the patronage of the Pallava kings of Kānchipuram, who began to rule in the 4th century AD and who financed the making of many temples and fine religious sculptures. Similar processes were taking place in the Deccan, Bengal, and other regions.

Sanskritization is a term that refers to a style of text that imitates the customs and manners of the Brahmans. But, although most sacred texts in Sanskrit were composed by Brahmans, many were also composed by lower-class authors. Likewise, although some sacred texts in vernacular languages were written by authors of lower castes, many others were written by Brahmans. In addition, because Sanskrit ceased to be spoken as a primary language soon after the Vedas were composed, it is likely that most of the thoughts underlying all subsequent Sanskrit literature were first thought in some other language. Yet Indologists tend to be Sanskritists, and Sanskritists tend to assume that all texts originated in Sanskrit. Indeed, even the counterbalancing tendency to acknowledge the flow from non-Sanskrit to Sanskrit sources has often misfired; far too often it is merely asserted that anything that appears in post-Vedic Hinduism and is not attested in the Vedas is "Dravidian," or, even worse, from the Indus civilization (about whose religion virtually nothing is known).

The issue is further clouded by the fact that, though Sanskrit texts tend to be written and vernacular traditions are primarily oral, there are important oral traditions in Sanskrit, too (including the traditions of the two great Sanskrit epics), and there are important manuscript traditions in some of the non-Sanskritic languages (such as Bengali and Tamil). Indeed, written and oral versions of the epics and *Purāṇa*s have been, from the very start, in constant symbiosis.

Little relevance, therefore, attaches to a distinction between "written" and "oral" traditions. A myth is essentially told or narrated, a process that is designated in Sanskrit by such words as *purāṇa* (ancient story) and *ākhyāna* (illustrative narrative). In the oldest source, the Rigveda, myths are not so much told as alluded to; it is in the later Vedic literature of the Brāhmaṇas that narratives are found, and these are often prejudiced by liturgical concerns of the authors.

Oral and written Hindu myths

The recitation of certain myths was prescribed for specific rituals. The epic *Mahābhārata* states that Vedic stories were narrated "in the pauses of the ritual," probably by Brahmans. The warrior class (Kṣatriyas) had their own mythographers in their *sūta*s (charioteers and panegyrists), who celebrated the feats of great rulers. These *sūta*s, who became popular narrators of myth and legend, had their own bardic repertoire, which soon was extended to higher mythology. They—and other wanderers who found ready audiences at sacrifices or places of pilgrimage—disseminated the lore.

Such narrators still continue to repeat and embroider their ancient stories of gods, sages, and kings. At an early stage their narratives were dramatized and gave rise to the Sanskrit theatre, in which epic mythic themes preponderate, and to the closely related dance, which survives in the now largely South Indian schools of bharata natya (traditional dance) and the kathakali (narrative dance) of Kerala. Thus, even in Sanskrit literature, oral performance was an essential component, which further facilitated the assimilation of oral vernacular elements.

When the Indo-Europeans, who spoke Sanskrit, an Indo-European language, entered India in around 1500 BC, most of the people they encountered spoke languages that belonged to a major non-Indo-European linguistic group

called Dravidian. These two language groups interacted from a very early period, and although the earliest preserved specimens of Sanskrit (themselves dating from a period long after the Indo-Aryan invasions) far antedate examples of any other languages, there is good reason to believe that the other languages also produced texts, although unrecorded, at a very early period. When the devotional aspect of Hinduism came into full flower, the vernacular traditions both in Dravidian language groups and in languages derived from Sanskrit began to record their texts and to have a more discernible influence upon the Sanskrit tradition.

Of the four primary Dravidian literatures—Tamil, Telugu, Kannada, and Malayalam—the oldest and best-known is Tamil. The earliest preserved Tamil literature, the so-called *Caṅkam* or *Śaṅgam* poetry anthologies, dates from the 1st century BC. These poems are classified by theme into *akam* ("interior," primarily love poetry) and *puṟam* ("exterior," primarily about war, the poverty of poets, and the deaths of kings). The bhakti movement has been traced to Tamil poetry, beginning with the poems of the devotees of Śiva called Nāyaṇārs and the devotees of Vishnu called Āḻvārs. The Nāyaṇārs, who date from about AD 800, composed intensely personal and devout hymns addressed to the local manifestations of Śiva.

The most famous Nāyaṇār lyricists are Appar, Sambandar, and Cuntarar, whose hymns are collected in the *Tevāram* (*c.* 11th century). More or less contemporary were their Vaishnava counterparts, the Āḻvārs Poykai, Pūtaṉ, Pēyār, and Tirumaṅkaiyāḻ-vār, and in the 8th century the poetess Āṇḍaḷ, Periyāḻvār, Kulacēkarar, Tiruppānāḻvār, and notably Nammālvār, who is held to be the greatest. The devotion of which they sing exemplifies the new bhakti movement that seeks a more direct contact between man and God, carried by a passionate love for the deity, who reciprocates by extending his grace to man. These saints also became the inspiration of theistic systematic religion: the Śaivas for the Śaivā-siddhānta, the Vaishnavas for Viśiṣṭādvaita. In Kannada the same movement was exemplified by Basava, whose *vācanams* ("sayings" or "talks") achieved great popularity. His religion, that of Vīraśaivism, was perhaps the most "protestant" of the bhakti religions.

Dravidian traditions

New Dravidian genres continued to evolve into the 17th and 18th centuries, when the Tamil Cittars (from the Sanskrit *siddhas*, "perfected ones"), who were eclectic mystics, composed poems noted for the power of their naturalistic diction. The Tamil sense and style of these poems belied the Sanskrit-derived title of their authors, a phenomenon that could stand as a symbol of the complex relationship between Dravidian and Sanskrit religious texts.

The main languages derived from Sanskrit are Bengali, Hindi (with its many dialects, of which Maithili is the oldest and Urdu, heavily influenced by Persian and Arabic and written in a Perso-Arabic script, is the most important), Punjabi, Gujarati, Marāṭhi, Oriya, Kashmiri, Sindhi, Assamese, Nepali, Rajasthani, and Sinhalese. Most of these languages began to develop literary traditions around AD 1000. The earliest texts in Hindi are those attributed to the 13th–14th-century Muslim poet Amīr Khosrow.

Hindi literature produced its own great religious lyricists beginning with the disciples of Rāmānanda (*c.* 1450), who was a follower of the philosopher Rāmānuja. Among them the most famous is Kabīr, whose bhakti was nonsectarian. Tulsīdās, apart from his *Rāmcaritmānas,* composed Rāmaite lyrics. Sūrdās (1483–1563), a follower of the Vallabha school of Vedānta, is famous for his *Sūrsāgar* ("Ocean of the Poems of Sūr"), a collection of poems based on the childhood of Krishna, following the account of the *Bhāgavata-Purāṇa.* In the Marāṭhi tradition, Nāmdev (*c.* 1300) celebrated Vishnu, particularly in his manifestation as Viṭṭhobā at the Pandharpur temple; and in the 17th century Tukārām, the greatest poet of this literature, sang of the god of love in numerous hymns.

A small sect, the Kabīrpanthīs, acknowledges Kabīr as its founder, but its importance is less than that of the vigorous new religion (Sikhism) founded by one of Kabīr's disciples, Nānak. In its final form, Sikhism contains elements taken from Islām (equality in the faith, opposition to iconolatry, extreme reverence for the sacred book) and probably also from Christianity (the Sikh baptism and communion meal), but its theology is still essentially Hindu.

Although the earliest Hindu text in Bengali is a mid-15th-century poem about Rādhā and Krishna, medieval texts in praise of gods and goddesses, known as *maṅgal-kāvyas,* must have existed in oral versions long before that. In later Bengal Vaishnavism, the emphasis shifts from service and surrender to mutual attachment and attraction between God (*i.e.,* Krishna) and humankind: God is said to yearn for the worshiper's identification with himself, which is his gift to the wholly purified devotee. The mystical and devotional possibilities of the Krishna legend are made subservient to religious practice; the divine sport and wonderful feats of this youthful hero are interpreted symbolically and allegorically. Thus, the highest fruition of bhakti is admission to the eternal sport of Krishna and his beloved Rādhā, whose sacred love story is explained as the mutual love between God and the human soul. Various gradations of bhakti are distinguished, such as awe, subservience, and parental affection. These are correlated with the persons of the Krishna legend; the highest and most intimate emotion is said to be the love of Rādhā and her girlfriends for Krishna.

A particularly rich tradition centred in Bengal concentrated on the love of Rādhā, who symbolizes the human soul, for Krishna, the supreme God. In this tradition are Caṇḍīdās and the Maithili poet Vidyāpati (*c.* 1400). The greatest single influence was Caitanya, who in the 16th century renewed Krishnaism. He left no writings but inspired many hagiographies, among the more important of which is the *Caitanya-caritāmṛta* ("Nectar of Caitanya's Life") by Krishna Dās (born 1517).

The work of Caitanya

Caitanya had a profound and continuing effect on the religious sentiments of his Bengali countrymen and propagated the community celebration (*saṃkīrtana*) of Krishna as the most powerful means of bringing about the proper bhakti attitude. Caitanya also introduced the worship of God, the director of man's senses, through the very activity of man's senses, which must be free from all egoism and completely filled with the intense desire (*preman*) for the satisfaction of the beloved (*i.e.,* Krishna).

The religious lyric continues in the so-called *pāda*s (verses); one of the greatest poets in this bhakti genre in which divine love is symbolized by human love is Govinda Dās (1537–1612). The songs of Rāmprasād Sen (1718–75) similarly honour Śakti as mother of the universe and are still in wide devotional use. The most famous religious lyrics in Gujarati are the poems of the saint Mīrā Bāī (1503–73), who wrote passionate love poems to Krishna, whom she regarded as her husband and lover.

The complex interaction between Sanskrit and non-Sanskrit religious classics may be seen in the development of the epics. Parts of the two great Sanskrit epics, the *Mahābhārata* and the *Rāmāyaṇa,* and many *Purāṇa*s (especially the *Bhāgavata-Purāṇa*) were translated into various vernaculars. Technically, these works were not literal translations, but free versions in which the authors placed their own emphases, different from the original and from one another. The oldest of the vernacular versions of the *Rāmāyaṇa* is the Tamil one of Kampaṉ (*c.* 12th century), a work suffused with devotion (bhakti) and of high literary distinction. Another famous translation in Tamil, written by Villiputturar, exists from the 18th century. A Telugu rendering was made by Ranganātha about 1300. In Bengali several translations were made, with some interesting and probably authentic variations from the "official" Rāma story by Vālmīki, the best-known one by Kṛttibās Ojhā (1450). Equally, if not more, famous is the Hindi version by Tulsīdās (*c.* 1550), entitled *Rāmcaritmānas* ("Holy Lake of the Acts of Rāma").

Regional language versions of the *Rāmāyaṇa* and *Mahābhārata*

The *Mahābhārata* was translated into Bengali about 1600 and into Telugu by Nannaya and Tikkana in the 13th century. The *Bhāgavata-Purāṇa,* which was translated frequently (*e.g.,* into Bengali by Maladhar Vasu, 1480), was popular both as a text and because it gave the canonical account of Krishna's life and especially his boyhood, which is the perennial inspiration of the bhakti poets.

In Marāṭhi the teacher Jñānadeva (also known as

Jñāneśvara; *c.* 1275–96) composed a commentary on the *Bhagavadgītā* that remains a classic in that literature. His work was continued by Eknāth (*c.* 1600), who also composed bhakti poetry. In the 16th century the Kannada poet Gadugu produced his own highly individual version of the *Mahābhārata.* In addition to the above literal or not-so-literal translations of the Sanskrit epics, the Tamils composed their own epics, notably Iḷaṅkō Aṭikaḷ's *Cilappatikāram* ("The Lay of the Anklet") and its sequel, the *Maṇimekhalai* ("Jeweled Girdle"). In Telugu there is the great Palnāḍu Epic; Rajasthani has an entire epic cycle about the hero Pabuji. The remaining vernaculars have produced many other works of the epic genre.

Myth and modernity

Hindu mythology in contemporary India. Much of the classical mythology persists today, and Hindus are exposed to it year-round. Meanwhile, the mass media have made their contributions: the type of motion picture called "mythological" is extremely popular, perpetuating the ancient stories to the village level, and so are "devotionals," in which an example of bhakti is illustrated. The radio regularly carries *bhajans* (devotional songs) and classical South Indian songs, the themes of which are often mythic. Every orthodox Hindu's home has at least one corner set aside as a domestic sanctuary where representations of a chosen deity are placed, and puja (worship) is done with prayers, hymns, flowers, and incense. Richer establishments set aside entire rooms as shrines. Mythic illustrations are favourites in Indian calendar art.

Mythology has adjusted itself effortlessly to modernity. The ashram of the 20th-century mystic and religious leader Śrī Aurobindo in Pondicherry, dedicated to the Mother Goddess (personified by this group as a single principle), is an extremely modern establishment complete with tennis courts. New temples are constructed with modern techniques; one temple in Vārānasi contains mirrors onto which are etched the entire *Rāmcaritmānas.* This same poem is the basis of the annual celebration of Rām Līlā (the play of Rāma) in northern India, in which the entire community participates. The Rāma story was evoked by Mahatma Gandhi when he set the Rām Rāj ("Kingdom of Rāma") as India's governmental ideal. On occasion, social protesters arm themselves with myth to make a point. For example, the personality of Karṇa, an antagonist in the *Mahābhārata* who is berated for his low birth, is extolled in intellectual circles as a truer champion than the aristocratic heroes. A Kannada-language play of the 1960s based on the life of King Yayāti enjoyed great popular and critical success. Anti-northern groups in Tamil Nadu revised the story of Rāma, whose expedition against the demon Rāvaṇa is believed by some to be the Aryan invasion of South India, by reversing it to abuse Rāma and to glorify Rāvaṇa.

On a popular level, people at temples and fairs are continually reacquainted with their mythological heritage by *paurāṇikas*, tellers of the ancient stories, heirs of the *sūtas* of 3,000 years ago, and no festival ground is complete without tents where the religious are reminded of their myths by pious speakers, modestly compensated by fees but richly rewarded by the honour in which they are held.

FOLK HINDUISM

Despite the impact of the West, the propaganda of modern reform movements, and the spread of education and secularist modernization, Hinduism has changed only slowly. For ordinary Hindus, religion primarily consists of the manual and verbal performance of rites to promote their private interests. The innumerable ceremonies, observances, fasts, feasts, pilgrimages, and visits to nearby temples constitute the essence of religion.

General characteristics of folk traditions. For millions, the main motive of religious practices is still the fear of ambivalent powerful beings. Most Hindus propitiate the meat-eating, sometimes benevolent but largely malevolent deities concerned with man's daily events, their ancestors or the founder of their community, and those various spirits that have no permanent residence and cause evil and misfortune. Hindus strive to escape the powers of the evil eye; to manipulate those spirits dwelling in wells, trees, stones, water, and ground; to counteract curses, witchcraft, plague, and cholera; and to worship village godlings who may give rain or a bountiful harvest. They make use of astrology, divination, and the reading of omens and auspicious moments. A large variety of purifications and ritual prohibitions, charms, and amulets to ward off any kind of misfortune (including bad luck in lawsuits and examinations) are, in the eyes of the majority, of greater importance than the atman-*brahman* doctrine. Even the hope of heaven or the fear of hell has little vogue in various regions, except among the higher castes.

It is difficult to draw a sharp line of distinction between popular Hinduism, the beliefs and practices of more or less Hinduized "external" groups, and Indian tribal religion. Many elements of tribal culture that in a particular region have not been adopted by those recognized as belonging to the Hindu fold are in fact similar to what has been adopted by Hindus in other areas. Tribal people and outcaste groups are, on the other hand, always willing to worship a few more gods or to imitate the rituals of lower-caste Hindus. Age-long processes of interpenetration and fusion have led to an adoption of many local and popular cults into general Hinduism—or, because it expressed itself mainly in Sanskrit, into the Great, or Sanskritic, tradition—and to the identification of regional gods with the great figures of the Hindu pantheon. Popular belief is integrated rather than discounted or discarded. This process is facilitated by a tendency toward the assimilation of local beliefs by pan-Indian Hinduism and by an unwillingness to deny gods and cults (the worship of a local river deity, for example, may be identified with that of the Ganges). The inheritors of the Little, or regional, traditions accepted, as a result of continual and complicated Hinduizing influences, vegetarianism, regular fasts, and food restrictions; they also began to object to the remarriage of widows, to observe Hindu festivals, to sing Hindu religious songs, to perform funeral and other ceremonious worship and, most importantly, to imbibe the ideas embodied in religious and mythological narratives. Thus, various tribal or outcaste groups have a religion with some affinities to a simple Śaivism without sacred books or regular liturgy.

Relation of folk Hinduism to Indian tribal religions

While many Hindus pursue their approach to the divine individually, corporate worship in families, villages, and sects is far more common in some castes. These groups exhibit the utmost variation in beliefs and practices. As a rule, each community practices only a small segment of the whole spectrum of religious behaviour as the expression of its own religious life. As to the relation between religion and social structure, there are in many communities differently structured systems, each with its own religious behaviour, in which their members may be involved. As members of a joint family, they take part in the domestic cult and ritually express family solidarity at such critical points as mourning or marriage; as members of a village community, they take part in its particular cult, which is a collective action of that community. Different castes, however, establish their own rituals in order to foster unity and to differentiate themselves from others. There is, on the other hand, ritual cooperation between different villages of the same region.

For many communities, spiritual reality is complex: while many women may address local spirits, family ancestors, and goddesses of disease, some of the men may embrace monotheistic ideas. The village's guardian spirit and the saint of a Muslim shrine may also be worshiped, Rāma's name is invoked in prayers, and the major deities are honoured chiefly during their periodic festivals. Marriage and other ceremonies combine ancient Sanskritic rites with popular and local features, and even members of the higher classes may accept the entire range of belief. In many regions, each caste has both general Hindu and "parochial" rituals and beliefs, but the proportions in which the two are found together vary from caste to caste and from locality to locality. The upper castes everywhere, however, have a certain amount of Sanskritic ritual in common; but even those who are more or less exclusively devoted to Śiva, for example, do not necessarily constitute a Śaiva community.

The bhakti movements have long influenced the religious feelings of their followers, and religious problems

are discussed by people of all professions and intellectual levels. Divine assistance is implored on every imaginable occasion; ancient Vedic rites have even been used as a defense against atomic danger.

Regional varieties of folk religion. In the hilly and mountainous regions of North India, Śaivism, aligned with Śāktism, is prevalent. The awe and mystery of the jungle and mountains are, there and elsewhere, personified as forest "Mothers" or mountain deities, represented by piles of stones or branches of trees to which every passerby contributes an offering. Mother Earth is a great goddess whose marriage (with the Earth god or the Sun) is festively celebrated and whose annual period of impurity is observed by a cessation of all agricultural activities. During the harvest season she is propitiated with wild orgies. In secluded parts of central India she is identified with Devī, a goddess mostly worshiped in North India. Very often, however, she shows herself in her malevolent form, as Mother Death (Mārī) or as Kālī. There are also many lower caste groups that have adopted a Vaishnava way of life either in order to raise their social status or to have a prospect of salvation.

Folk goddesses

In the east and northeast, where, broadly speaking, Śāktism is dominant, though Vaishnavism is also common, popular belief has modified the transmigration doctrine by the assumption that the soul of the deceased reappears in a child born in the same family within a year after the person's death. Among the female deities, there are tutelary goddesses of young children and women in childbed: Ṣaṣṭhī, "the Sixth," is worshiped on the sixth day after birth and is represented by a compost pile of cow dung or earth that is placed in the birth room; and Caṇḍī, a form of the goddess Durgā, lives in trees and is propitiated by lumps of earth. The snake goddess Manasā is personified in a plant of the same name or in a stone carved into the shape of a female seated on a snake; a day in the rainy season, when reptiles are most dangerous, is devoted to her priestless and unpretentious worship. In literary works she is eulogized as the Great Mother who is expected to give a prosperous journey through life and, to a certain extent, is also Sanskritized by being identified with epic snake demons. Another example of fusion of general and local Hindu institutions is the conviction that ghosts and demons are warded off by performing a ceremony in honour of the deceased at Gayā in modern Bihār state.

In many regions—especially in western India, where Vaishnavism is dominant—believers admit that virtue will improve their lot in a subsequent existence, but they do not seem to strive for final union with the Supreme. Here, and elsewhere, a workaday religion meant to meet the requirements of everyday life exists alongside a higher religion understood only by the Brahmans, who are called on to officiate on important occasions. In order to discover the divine will, exorcists and mediums, possessed by mother goddesses and submitting themselves to self-torture, are called upon to prophesy about future events. In these regions the worship of snakes is more prominent, and some temples are even dedicated to them. Practices based on the belief in scapegoats, ritual nudity, and black magic are also widespread.

The whole of peninsular India is mainly devoted to Śaivism, devotional forms of Vaishnavism, and the worship of the goddess in her many forms. A striking feature in the religion of South India is the propitiation of usually local female village deities of varied and ambivalent character, to whom in almost every settlement a simple shrine or other sacred place is dedicated. These deities are thought to be particularly competent for dealing with the facts of village life, such as diseases of the inhabitants and their cattle. Special cholera and smallpox goddesses are the subject of elaborate stories. In a few cases—*e.g.*, that of Māriyammā, the smallpox goddess of South India—such a goddess is known to a large region. These mothers, from whom all good and bad luck emanate, are almost universally worshiped with animal sacrifices, and the priestly ministrants (*pūjārī*) officiating in their cult belong to the non-Brahman groups. The goddesses may be represented by various symbols (stone pillars, sticks, clay figures) that need not be permanent. Most of their shrines are simple, small brick buildings or rough stone platforms under a tree. Offerings of rice, fruit, and flowers may be made every day or on fixed days; although there often is a fixed annual festival, it is not uniformly celebrated and no calendar of festivals is established. An exception to this is the male deity Aiyanār, who in the countryside of Tamil Nādu state is worshiped as the watchman and patron of the villages but also is implored to grant children and other blessings. He is a vegetarian and therefore ranks as socially superior to the female village goddess with whom he has entered into a complementary relation. Aiyanār is worshiped either as a village deity or in a temple dedicated to Śiva, where he is given the rank of a son of that god. In these Śiva temples he is legitimated by higher Hinduism and fulfills the function of a double of Śiva representing "All-India" or general Hinduism in the village, which does not regard him as an outsider. Śiva himself is also worshiped and given a consort, who, though considered a manifestation of Durgā, has various names according to the tradition of temple or village.

Aiyanār

Sacred snakes, especially cobras, are also given offerings—partly to avert danger from these reptiles, partly to propitiate them with the aim of obtaining rain, fertility, or children; to that end women worship snake stones (*nāgalkals*) or erect stone figures of cobras. Every joint family of the Coorgs in Karnātaka state and most other peoples have a snake deity of their own that is said to embody their welfare. Here and there, Brahmans officiate in this cult, which usually takes place in small sanctuaries in private gardens. Although also known in other parts of India, the methods of exorcising evil spirits known as devil-dancing are most fully developed in South India. The notorious hook-swinging festival, Caḍak-pūjā, held for propitiatory purposes in cases of famine or other calamities—a man was suspended by hooks at the end of a long pole and swung around—though strictly prohibited, survived to the 20th century. Greater festivals are, generally speaking, either celebrated at the chief agricultural seasons or connected with the expulsion of malign powers.

Folk and tribal myths. There is a great diversity in folk mythology throughout the entire Indo-Pakistan subcontinent, but these myths have neither been fully collected nor systematically studied. Among locally important deities, Manasā, a snake goddess, worshiped in Assam and Bengal to ward off snakebites and secure prosperity, has an enormous mythology of her own. In South India one finds popular cobra cults with a variety of myths and lore. In Mahārāshtra, a form of Vishnu, known as Viṭṭhal or Viṭṭhobā, has also spawned a rich mythology.

The sources of folk and tribal mythology are vernacular literature, oral tradition, folklore, and folk and tribal arts. Folk mythology derives from the most ancient times and has influenced both Vedic and classical mythology. Classical mythology became what it is by continuously assimilating myths not previously known or accepted, so that the line between classical and folk (and tribal) mythology is apt to be arbitrary. The Great (Sanskrit) tradition of classical mythology (as contrasted with the local, or Little, traditions of folk mythology) may include within its scheme a god who continues to have an independent existence on a folk level. Sacred manifestations of purely local interest are associated with the higher mythology by becoming a Little manifestation of a Great god, such as the footstep of Rāma and the bathing place of Sītā (Rāma's wife). Conversely, an incident of the Great tradition may be adopted and adapted on the folk level. For example, the local Mahārāshtrian god Viṭṭhobā is identified with a manifestation of Vishnu and thus assured a place in the Great tradition; on the other hand, in North India, the widely celebrated festival of Navarātrī ("Nine Nights") is associated with the village goddess Naurthā.

The concepts of avatar and *vāhana*

Certain concepts that evolved in Purāṇic mythology have facilitated the absorption of folk elements; two of these should be singled out: avatar (*avatāra,* incarnation), and *vāhana* (vehicle).

The concept of avatar (literally "descent") issues from the belief that in times of trouble a god, notably Vishnu, incarnates himself as a man or hero to set matters right. Such a concept provides the opportunity for identifying a

local deity (like Viṭṭhobā, above) with an all-Indian god like Vishnu. The concept may also extend to the worship of very local hierophanies (manifestations of the sacred; *e.g.,* South Indian Vaishnavism accepts "icon-incarnation" [*arcāvatāra*], in which Vishnu "descends" in a local icon).

According to the concept of a *vāhana* (literally "mount"), every god has an entourage of his own, which includes a favourite riding animal; this facilitates many folk associations. Vishnu's mount is the bird Garuḍa, an old solar symbol; Śiva's is the bull Nandi, whose worship may go back to the ancient Harappan civilization. There are other mythological patterns, such as adoption in a family; thus the folk god Gaṇeśa, an elephant-headed god, is made the son of Śiva, as is Kumāra Kārttikeya, the war god, who arose from the South Indian war god Muruganṉ. Hanumān, the monkey god, becomes an all-Indian god as a helper of Rāma, who is an avatar of Vishnu.

Other spirits and godlings of folk provenance are not absorbed to the same degree and thus retain their folk character. Important are the snakes (nagas), to which great power is attributed; the yakshas, koboldlike keepers of wealth, whose king is Kubera; *vetāla*s, ghoulish pranksters who haunt corpses; and spirits of the restless dead (bhuts, pretas), who must be warded off. Though the existence of these spirits is fully recognized in the classical mythology, they operate primarily on the folk level.

It is therefore difficult to find folk myths and cults that are not in one form or another elevated into the Sanskrit tradition by a specific association with a major god. Only where there has been no appreciable cultural contact between Indian tribal people and Hindus (or "Hinduized" folk groups) can a significant distinction be drawn between classical Hindu mythology and "folk" mythology. The myths of the tribes of Chota Nāgpur (Bihār state), the Santāl (West Bengal and Bihār states), the Toda in the Nīlgiri Hills (Tamil Nādu state), and others are examples of such mythology. Much religious material lies hidden in folktales.

(J.A.B.v.B./E.C.D./A.L.B./W.Do./B.K.S./Ed.)

Rituals, social practices, and institutions

SACRIFICE AND WORSHIP

Although the Vedic fire rituals were largely replaced in Purāṇic and modern Hinduism by image worship and other forms of devotionalism, many Hindu rites can still be traced back to Vedism. Certain royal sacrifices—such as the *rājasūya,* or consecration ritual, and the horse sacrifice (*aśvamedha*)—remained popular with Hindu kings until very recently. Other large-scale Vedic sacrifices (*śrauta*) have been regularly maintained from ancient times to the present by certain families and groups of Brahmans. By and large, however, the surviving rituals from the Vedic period tend to be most clearly observed at the level of the domestic (*gṛhya*) ritual.

Domestic rites. The Vedic householder was expected to maintain a domestic fire into which he made his offerings. Normally he did this himself, but in many cases he employed a Brahman officiant. In the course of time, the family priest was given a large part in these ceremonies, so that most Hindus have employed Brahmans for the administration of the "sacraments" (samskara). The samskaras include all important life-cycle events, from conception to cremation, and are the main constituents of the domestic ritual.

Rites of passage

The sacraments are transitional rites intended to make a person fit for a certain purpose or for the next stage in life, by removing taints (sins) or by generating fresh qualities. If the blemishes incurred in this or a previous life are not removed, the person is impure and will acquire no reward for any ritual acts. The sacraments, while sanctifying critical moments, are therefore deemed necessary for unfolding a person's latent capacities for development. Śūdras are allowed to perform some samskaras if they do not require the use of Vedic mantras.

Samskaras: rites of passage. In antiquity there was a great divergence of opinion about the number of rites of passage, but in later times 16 were regarded as the most important. The impregnation rite, consecrating the supposed time of conception, consists of a ritual meal of pounded rice (mixed "with various other things according to whether the married man desires a fair, brown, or dark son; a learned son; or a learned daughter"), an offering of rice boiled in milk, the sprinkling of the woman, and intercourse; all acts are also accompanied by mantras. In the third month of pregnancy the rite called *puṃsavana* (begetting of a son) follows. The birth is itself the subject of elaborate ceremonies, the main features of which are an oblation of ghee (clarified butter) cast into the fire; the introduction of a pellet of honey and ghee into the newborn child's mouth, which according to many authorities is an act intended to produce mental and bodily strength; the murmuring of mantras for the sake of a long life; and rites to counteract inauspicious influences. There is much divergence of opinion as to the time of the name-giving ceremony; in addition to the personal name, there is often another one that should be kept secret for fear of sinister designs against the child.

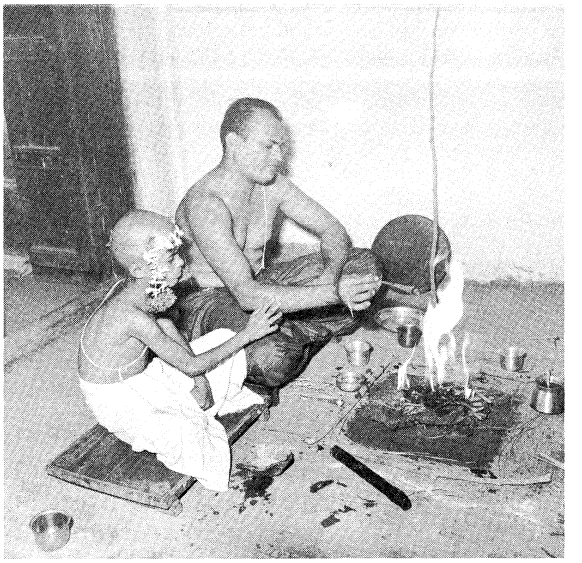
Rites associated with the newborn

In modern times most samskaras (with the exceptions of impregnation, initiation, and marriage) have in many areas fallen into disuse or are performed in an abridged or simplified form without Vedic mantras or a priest. This tendency was encouraged by the accommodating spirit of the Brahmans, who allowed their clients easy atonements for the nonobservance of rites. The important *upanayana* initiation is held when a boy is between the ages of eight and 12 and marks his entry into the community of the three higher classes of society. In this rite he becomes a "twice-born one," or *dvija.* Traditionally, this was also the beginning of a long period of Veda study and education in the house under the guidance of a teacher (guru). In modern practice, the haircutting ceremony—formerly performed in a boy's third year—and the initiation are usually performed on the same day, the homecoming ceremony at the end of the period of study being little more than a formality.

Marriage customs

Wedding ceremonies, the most important of all, have not only remained elaborate—and often very expensive—but have also incorporated various elements—among others, propitiations and expiations—that are not indicated in the oldest sources. Already in ancient times there existed great divergences in accordance with local customs or family or caste traditions. However, the following practices are usually considered essential. The date is fixed only after careful astrological calculation; the bridegroom is conducted to the home of his future parents-in-law, who receive him as an honoured guest; there are offerings of roasted grain into the fire; the bridegroom has to take hold of the bride's hand; he conducts her around the sacrificial fire; seven steps are taken by bride and bridegroom to solemnize the irrevocability of the unity; both are, in procession, con-

McKim Marriott



A young boy performs his first puja after initiation into the community of the "twice-born."

ducted to their new home, which the bride enters without touching the threshold.

Of eight forms of marriage recognized by the ancient authorities, two have remained in vogue: the simple gift of a girl and the legalization of the alliance by means of a marriage gift paid to the bride's family. In the Vedic period, girls do not seem to have married before they had reached maturity. Child marriage and the condemnation of the remarriage of widows, especially among the higher classes, became customary later and have gradually, since the mid-19th century, lost their stringency.

Death rites and customs

The traditional funeral method is cremation (a family affair), burial being reserved for those who have not been sufficiently purified by samskaras (*i.e.,* children) and those who no longer need the ritual fire to be conveyed to the hereafter, such as ascetics who have renounced all earthly concerns. An important and meritorious complement of the funeral offices is the sraddha ceremony, in which food is offered to Brahmans for the benefit of the deceased. Many people are still solicitous to perform this rite at least once a year even when they no longer engage in any of the five obligatory daily offerings.

Daily offerings. There are five obligatory offerings: (1) offerings to the gods (food taken from the meal); (2) a cursory offering (*bali*) made to "all beings"; (3) a libation of water mixed with sesame offered to the spirits of the deceased; (4) hospitality; and (5) recitation of the Veda. Although some traditions prescribe a definite ritual in which these five "sacrifices" are performed, in most cases the five daily offerings are merely a way of speaking about one's religious obligations in general.

Other private rites. The morning and evening adorations (*sandhyā*), being a very important duty of the traditional householder, are mainly Vedic in character, but they have, by the addition of Purāṇic and Tantric elements, become lengthy rituals. If not shortened, the morning ceremonies consist of self-purification, bathing, prayers, and recitation of mantras, especially the Gāyatrī mantra (Rigveda 3.62.10), a prayer for spiritual stimulation addressed to the Sun. The accompanying ritual includes (1) the application of marks on the forehead, characterizing the adherents of a particular religious community, (2) the presentation of offerings (water, flowers) to the Sun, and (3) meditative concentration. There are Śaiva and Vaishnava variants, and some elements are optional. The observance of the daily obligations, including the care of bodily purity and professional duties, leads to mundane reward and helps to preserve the state of sanctity required to enter into contact with the divine.

Gāyatrī mantra

Temple worship. Image worship in sectarian Hinduism takes place both in small shrines in each house and in the temple. Many Hindu authorities claim that regular temple worship to one of the deities of the devotional cults procures the same results for the worshiper as did the performance of one of the great Vedic sacrifices, and one who provides the patronage for the construction of a temple is called a "sacrificer" (*yajamāna*).

Temples. The erection of a temple, which belongs to whoever paid for it or to the community that occupies it, is a meritorious deed recommended to anyone desirous of heavenly reward. The choice of a site, which should be serene and lovely, is determined by astrology and divination as well as by its location with respect to human dwellings; for example, a sanctuary of a benevolent deity should face the village. The construction of a temple is, because of its symbolic value, described in great detail. There is much diversity in size and artistic value, ranging from small village shrines with simple statuettes to great temple-cities whose boundary walls, pierced by monumental gates (gopura), enclose various buildings, courtyards, pools for ceremonial bathing, and sometimes even schools, hospitals, and monasteries. From the point of view of construction there is no striking difference between Śaiva and Vaishnava sanctuaries, which are easily recognizable by the image or symbols in the centre, the images on the walls, the symbol fixed on the finial (crowning ornament) of the top, and Śiva's bull, Nandi, or Vishnu's bird, Garuḍa (the theriomorphic duplicate manifestations of each god's nature), in front of the entrance. Services, which may be held by any qualified member of the community, are neither collective nor carried out at fixed times. Those present experience, as spectators, the fortifying and beneficial influence radiating from the sacred acts. Sometimes worshipers assemble to meditate, to take part in chanting, or to listen to an exposition of doctrine. The puja (worship) performed in public "for the well-being of the world" is, though sometimes more elaborate, largely identical with that executed for personal interest. There are, on the other hand, many regional differences, and even significant variations within the same community.

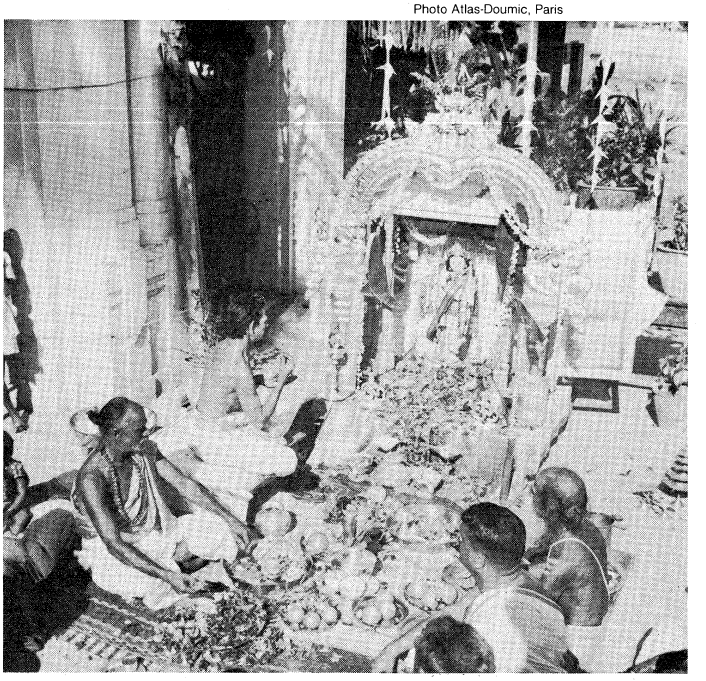
Puja. Hindu worship (puja) consists essentially of an invocation, a reception, and the entertainment of God as a royal guest. It normally consists of 16 "attendances" (*upacāra*): invocation by which the omnipresent God is invited to direct his attention to the particular worship; the offering of a seat, water (for washing the feet, for washing the hands, and for rinsing the mouth), a bath, a garment, a sacred thread, perfumes, flowers, incense, a lamp, food, and homage; and a circumambulation of the image and dismissal by God.

The Pāñcarātra Vaishnavas in South India introduced the songs of the Dravidian poets into their temple cult and regard these poets and their great teachers as incarnations of God, even to the point of worshiping their images. The Śaivas also have songs of their own but were, generally speaking, more open to Tantric elements and to the admission in their cult of dances executed by dancing girls. In both religious groups, some communities cling to the traditional Sanskrit mantras while others also use other languages.

The significance of the temple door

The first phase of worship is the reverential opening of the temple door and the adoration of the powers presiding over it: according to the Vaikhānasa Vaishnavas, the symbolic opening of heaven; and to the Śaivas, an act to secure the building's protection. The divine powers whose images are carved in the doorjambs promote the process of transmutation without which man cannot even enter into the presence of God, whose image is established in the cella (*garbhagṛha*). This image is honoured with gifts, notably flowers, fruit, and perfumes. Small portions of the consecrated food (*prasāda*) are given to visiting worshipers. The offering into the fire (*homa*) of Vedic origin has been retained in nearly all extended puja ceremonies. The main purpose of the rites is the meditative identification of the worshiper with the divine Presence; the enactment, in a gradual process of development, of the realization of the union of the worshiper's soul and God. The Vaikhānasas distinguish between the transcen-

Photo Atlas-Doumic, Paris



A temple puja to Krishna.

dent and unanalyzable Brahman and its immanent and analyzable aspect and invoke God to descend out of compassion from the immovable image—the permanent "seat" of the former—into a movable cult image in which he converses with the world, represented by the worshiper. Those denominations (both Śrīvaiṣṇavas and Śaivas) that adopted Tantric practices believe that God comes, during these ceremonies, also out of the worshiper's heart or that the worshiper's soul leaves his body to reach God's feet in heaven, to descend from there in a new body that is meditatively created.

A remarkable rite of yogic-Tantric origin, also used in other ritual contexts, is the transmutation of water into the elixir of life and immortality (*amṛta*), the essential element of which is drawn from the spot between the worshiper's eyebrows, regarded as the seat of Śiva's highest aspect.

Śaivas transform themselves into Śiva by means of complicated preparatory rites, because, they say, "Śiva alone can worship Śiva." Some authorities also enjoin a mental worship and sacrifice, without which "exterior" rites are rendered senseless. The merit of the performances is often said to be entrusted to God's keeping for the sake of the worshiper. Many Vaishnavas emphasize that puja is meant to propitiate God disinterestedly.

Śaiva rites. Ascetic tendencies were much in evidence among the Pāśupatas, the oldest Śaiva tradition in North India, the last adherents of which now live in Nepal. Pāśupatas often gave offense because of their customs and ritual practices. Their yoga, consisting of a constant meditative contact with God in solitude, required that they frequent burning places for cremated bodies. More extreme groups carried human skulls (hence the name Kāpālikas, from *kapāla,* "skull") which they used as bowls for liquor into which they projected and worshiped Śiva as Kāpālika, "the Skull Bearer," or Bhairava, "the Frightful One," and then drank to become intoxicated. Their belief was that an ostentatious indifference to anything worldly was the best method of severing the ties of samsara.

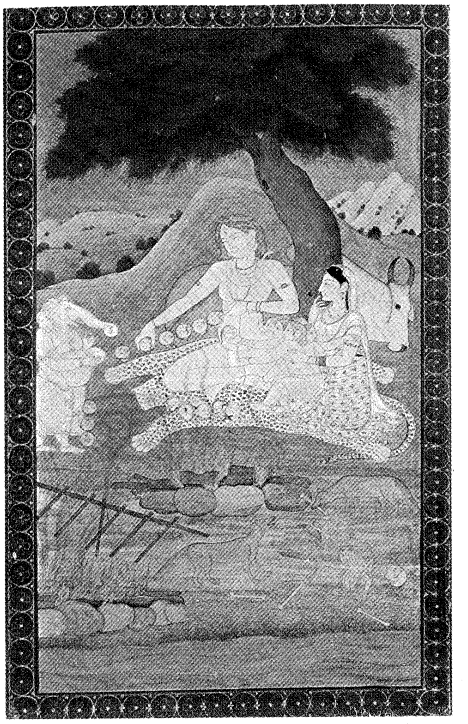
Vīraśaivas (Liṅgāyats)

The view and way of life peculiar to the Vīraśaivas, or Liṅgāyats (lingam-bearers), in southwestern India is mainly characterized by a deviation from some common Hindu traditions and institutions such as sacrificial rites, temple worship, pilgrimages, child marriages, and inequality of the sexes. Initiation (*dīkṣā*) is, on the other hand, an obligation laid on every member of the community. The spiritual power of the guru is bestowed upon the newborn and converts, who receive the eightfold shield, which protects devotees from ignorance of the supremacy of God and guides them to final beatitude, and the lingam (phallic symbol). The miniature lingam, the centre and basis of all their religious practices and observances, which they always bear on their body, is God himself concretely represented. Worship is due it twice or three times a day. When a Liṅgāyat "is absorbed into the lingam" (*i.e.,* dies), his body is not cremated, as is customary in Hinduism, but is interred, like ascetics of other groups. Those Liṅgāyats who have reached a certain level of holiness are supposed to die in the state of emancipation.

Varieties of Śaivism

Śaivism, though inclined in doctrinal matters to adoptive inclusivism, inculcates some fundamental lines of conduct: one should worship one's spiritual preceptor (guru) as God himself, follow his path, consider him to be present in oneself, and dissociate oneself from all opinions and practices that are incompatible with the Śaiva creed. Yet some of Śiva's devotees also worship other gods, and the "Śivaization" of various ancient traditions is sometimes rather superficial. Like many other Indian religions, the Śaiva-siddhānta has developed an elaborate system of ethical philosophy, primarily with a view to preparing the way for those who aspire to liberation. Because dharma leads to happiness, there is no distinction between sacred and secular duties. All deeds are performed as services to God and with the conviction that all life is sacred and God-centred. A devout way of living and a nonemotional mysticism are thus much recommended. Kashmir Śaivism developed the practice of a simple method of salvation: by the recognition (*pratyabhijñā*)—direct, spontaneous, technique-free, but full of bhakti—of one's identity with God.

Vaishnava rites. The day of the faithful Śrīvaiṣṇava



Śiva and his family at the burning ground. Pārvatī, Śiva's wife, holds Skanda while watching Gaṇeśa (left) and Śiva string together the skulls of the dead. The bull, Nandi, rests behind the tree. Kangra painting, 18th century; in the Victoria and Albert Museum, London.

By courtesy of the Victoria and Albert Museum, London; photograph, A.C. Cooper

Brahman is usually devoted to five pursuits: purificatory rites, collecting the requisites for worship, acts of worship, study and contemplation of the meaning of the sacred books, and meditative concentration on the Lord's image. Lifelong obligations include the performance of sacrifices and other rites, restraint of the senses, fasting and soberness, worship, recitation of the scriptures, and visits to sacred places. In addition, to those who aspire to liberation, Rāmānuja recommends concentration on God, a virtuous way of living, and insensibility to luck and misfortune. According to Madhva (*c.* 1199–*c.* 1278), a faithful observance of all regulations of daily conduct—including bathing, breath control, etc.—will contribute to eventual success in the quest for liberation. Devout Vaishnavas are inclined to emphasize God's omnipotence and the far-reaching effects of his grace. They attach much value to the repeated murmuring of his name or sacred formulas (*japa*) and to the praise and commemoration of his deeds as a means of self-realization and of unification with his essence. Special stress is laid on ahimsa as a virtue.

SACRED TIMES AND PLACES

Festivals. Hindu festivals are combinations of religious ceremonies, semi-ritual spectacles, worship, prayer, lustrations, processions (to set something sacred in motion and to extend its power throughout a certain region), music, dances (which by their rhythm have a compelling force), magical acts—participants throw fertilizing water or, during the Holī festival, coloured powder at each other—eating, drinking, lovemaking, licentiousness, feeding the poor, and other activities of a religious or traditional character. The original functions of these activities are clear from ancient literature and anthropological research: they are intended to purify, avert malicious influences, renew society, bridge over critical moments, and stimulate or resuscitate the vital powers of nature (hence the term *utsava,* meaning both the generation of power and a festival). Because such festivals relate to the cyclical life of nature, they are supposed to prevent it from stagnating. These cyclic festivals—which may last for many days—continue to be celebrated throughout India.

Such festivals refresh the mood of the participants, further the consciousness of their own power, and help to compensate for their sensations of fear and inferiority concerning the unknown forces of nature. Such mixtures of worship and pleasure require the participation of the entire community and create harmony among its members, even if not all participants are now aware of the original character of the festival. There are also innumerable festivities in honour of specific gods, celebrated by individual temples, villages, and religious communities.

An important festival, formerly celebrating Kāma, the god of sexual desire, survives in the Holī, a saturnalia connected with the spring equinox and in western India with the wheat harvest. The lower classes observe it in its boisterous and licentious form. There are local variants: among the Marāṭhās, heroes who died on the battlefield are "danced" by their descendants, sword in hand, until they believe themselves possessed by the spirits of the heroes. In Bengal, swings are made for Krishna; in other regions a bonfire is also essential. The mythical tradition that accounts for the festival describes how young Prahlāda, in spite of his demonic father's opposition, persisted in worshiping Vishnu and was carried into the fire by the female demon Holikā, the embodiment of evil, who herself was believed to be immune to the ravages of fire. Through Vishnu's intervention, Prahlāda emerged unharmed, while Holikā was burned to ashes. The bonfires are intended to commemorate this event or rather to reiterate the triumph of virtue and religion over evil and sacrilege. This explains why objects representing the sickness and impurities of the past year—the new year begins immediately after Holī—are thrown into the bonfire, and it is considered inauspicious not to look at it. Moreover, people pay or forgive debts, reconcile quarrels, and try to rid themselves of the evils, conflicts, and impurities that have accumulated during the preceding months, translating the central conception of the festival into a justification for dealing anew with continuing situations in their lives. The New Year festival, according to another Indian calendar, Dīwālī, though celebrated by all classes of society, is traditionally believed to have been given by Vishnu to the Vaiśyas (traders, etc.); it takes place in October, with worship and ceremonial lights in honour of Lakṣmī, the goddess of wealth and good fortune; fireworks to chase away the spirits of the deceased; and gambling, an old ritual custom intended to secure luck for the coming year. The nine-day Durgā festival, or Navarātrī, is, especially in Bengal, splendid homage to Śakti, and in South India, a celebration of Rāma's victory over Rāvaṇa.

New Year festivals

Pilgrimages and fairs. Like processions, pilgrimages (*tīrthayātrā*) to holy rivers (*tīrtha*) and other places were already known in Vedic and epic times and are even now one of the most remarkable aspects of Indian religious life. Many sections of the *Purāṇas* eulogize temples and the sacredness of places situated in beautiful scenery or wild solitude (especially the Himalayas). The whole of India, and especially Kurukshetra (presumed to be the scene of the great war portrayed in the *Mahābhārata*) in the northwest, is considered holy ground that offers everyone the opportunity to reach emancipation. The number of places of pilgrimage of regional significance amounts to many hundreds, but some of them (Ayodhyā, Mathura, Hardwār, Vārānasi [Benares], Kānchipuran, Ujjain, and Dwārka) have for many centuries possessed exceptional holiness. The reason for such sanctity derives from their location on the bank of a holy river, especially of the Ganges, from their connection with legendary figures of antiquity who are said to have lived there, or from the local legend of a manifestation of a god. Many places are sacred to a specific god; the district of Mathura, for example, encompasses many places of pilgrimage connected with the Krishna legends. Visits to holy places may bestow special benefits upon pilgrims; temples or ponds dedicated to Sūrya (the Sun) are visited in order to recover from leprosy, other places to escape from astrological threats. Pilgrimages to Gayā (Bihār state)—where visitors are escorted around the sacred centres by Brahman temple priests who maintain certain ritual connections with their clients—are undertaken for the sake of the welfare of deceased ancestors. In most cases, however, the devotee hopes for worldly rewards (health, wealth, children) or for spiritual rewards such as deliverance from sin or pollution, preservation of religious merit, rebirth in a heaven, or even emancipation. The last prospect is held out to those who, when death is near, travel to Vārānasi to die near the Ganges.

On special occasions, be they auspicious or, like a solar eclipse, inauspicious, the devout crowds increase enormously. Most important shrines also organize gatherings (melas), that are partly fairs, partly religious demonstrations. These journeys, which are undertaken by individuals or groups in order to discharge a vow or to please a god, confirm the devotees in their faith, provide them with an opportunity for spiritual retreat, or bring their inner life nearer to a state of perfection. They have contributed much to the spread of religious ideas and the cultural unification of India.

RITUAL AND SOCIAL STATUS

Some observers have claimed that Hinduism is as much a way of social life as it is a religion. The caste system, which has organized Indian society for many millennia, is thoroughly legitimated by and intertwined with Hindu religious doctrine and practice. Four social classes, or varnas—Brahmans, Kṣatriyas, Vaiśyas, and Śūdras—provide the simplified structure for the enormously complicated system of thousands of castes and subcastes within Indian society. Although it is not certain whether a society limited to four classes was ever more than a theoretical ideal, there is a sense in which they map out socioreligious reality. Such is evident from the Puruṣa hymn (Rigveda 10.90), in which the statement that the Brahman was the Puruṣa's mouth, the nobleman (Kṣatriya) his arms, the Vaiśya his thighs, and the Śūdra his feet, gives an idea of their functions and mutual relations.

The Brahmans, whatever their worldly avocations, claim to be by virtue of their birth a perpetual incarnation of the dharma, guardians and dispensers of divine power, entitled to teach the Veda, sacrificing for others and accepting gifts and subsistence; the term alms is misleading, and the *dakṣiṇā* offered at the end of a rite to a Brahman officiant is not a fee but an oblation through which the rite is made complete. Brahmans are held to be the highest of all human beings because of their preeminence, the superiority of their origin, their sanctification through the samskaras (rites of passage), and their observance of restrictive rules. The main duty of the nobility (the Kṣatriyas) is to protect the people, that of the commoners (the Vaiśyas) to tend cattle, to trade, and to cultivate land. Even if a king (theoretically of Kṣatriya descent) was not of noble descent, such an upholder of dharma was clothed with divine

Paul Popper Ltd.



Pilgrims bathing in the Ganges River at Hardwār, India.

authority. He was consecrated by means of a complex and highly significant ritual; he was Indra and other gods (deva) incarnate. The emblems or paraphernalia of his office represent sovereign authority; the white umbrella of state is, for example, the residence of Śrī-Lakṣmī, the goddess of fortune. All three higher classes, claiming Aryan descent, had to sacrifice and to study the Veda, although the responsibilities of the Vaiśyas in sacred matters were less demanding.

While this tripartition seems, in the main, to have been inherited from Indo-European times, the fourth class (the Śūdras), whose sole duty it was "to serve meekly" (*Mānava Dharmaśāstra* 1.91) the other classes, are partly descended from the subjugated non-Aryans, a fact that accounts for their many disabilities and exclusion from religious status. According to Hindu tradition, the Veda should not be studied in their presence, but they may listen to the recitation of epics and *Purāṇa*s. They are permitted to perform the five main acts of worship (without Vedic mantras) and undertake observances, but even today they maintain various ceremonies of their own, carried out without Brahmanic assistance. Yet a distinction is often made among Śūdras. Some are purer and have a more correct behaviour and way of living than others, the former tending to assimilate with higher castes, the latter to rank with the lowest in the social scale, who, often called *cāṇḍāla*s, were at an early date sweepers, bearers of corpses, or charged with other impure occupations. Ritual purity was indeed an important criterion; impure conduct and neglect of Veda study and the rules regarding forbidden food might suffice to stigmatize a "twice-born man" as a Śūdra. On the other hand, in later times the trend of many communities has been toward integrating all Śūdras into the Brahmanic system. The Brahmans, who have far into modern times remained, on the whole, a respected, traditional, and sometimes intellectual upper class, were generally (until the 1930s) much in demand because of their knowledge of rites and traditions. Although Kṣatriya rank is claimed by many whose title is one of function or creation rather than of inheritance, this class is now rare in many regions. Moreover, for a considerable time none of the four varnas represented anything other than a series of hierarchically arranged groups of castes.

Castes. The origin of the caste system is not known with certainty. Hindus account for the proliferation of the castes (*jāti*s, literally "births") by the subdividing of the four classes, or varnas, due to intermarriage (which is prohibited in Hindu works on dharma). Modern theorists, however, tend to assume that castes arose from differences in family ritual practices, racial distinctions, and occupational differentiation and specialization. Many modern scholars also doubt whether the simple varna system was ever more than a theoretical socioreligious ideal and have emphasized that the highly complex division of Hindu society into nearly 3,000 castes and subcastes was probably in place even in ancient times.

In general, a caste is an endogamous hereditary group of families, bearing a common name; often claiming a common descent; as a rule professing to follow the same hereditary calling; clinging to the same customs, especially regarding purity, meals, and marriages; and often further divided into smaller endogamous circles. Moreover, tribes, guilds, or religious communities characterized by particular customs—for example, the Lingāyats—could easily be regarded as castes. The status of castes varies in different localities. Although social mobility is possible, the mutual relationship of castes is hierarchically determined: local Brahman groups occupy the highest place, and differences in ritual purity are the main criteria of position in the hierarchy. Most impure are the untouchables, or, to use modern names, the exterior or scheduled castes, which, however, have among themselves numerous divisions, each of which regards itself as superior to others.

Traditional Hindus are inclined to emphasize that the ritual impurity and "untouchability" inherent in these groups does not essentially differ from that temporarily proper to mourners or menstruating women. This, and the fact that some exterior group or other might rise in estimation and become an interior one, or that individual outcastes might be well-to-do, does not alter the fact that the spirit of exclusiveness was in the course of time carried to extremes. The scheduled castes were subjected to various socioreligious disabilities before mitigating tendencies helped bring about reform; after independence, social discrimination was prohibited, and the practice of untouchability was made a punishable offense (it was not abolished, however). Scheduled castes were barred from the use of temples and other religious institutions and from public schools. These groups also had many disabilities in relations with private persons. From the traditional Hindu point of view, this social system is the necessary complement of the principles of dharma, karma, and samsara. Corresponding to hells and heavenly regions in the hereafter, the castes are the mundane, social frame within which karma is manifested. A low social status is the inevitable result of sins in a former life but can, by virtue and merit, be followed by a better position in the next existence.

Religious orders and holy men. Those members of the various denominations who abandon all worldly attachment enter an "inner circle" or "order" that, seeking a life of devotion, adopts or develops particular vows and observances, a common cult, and some form of initiation.

Initiation. Generally speaking, Hindus are free to join an order or inner circle, and once they have joined it they must submit to its rites and way of living. The initiation (*dīkṣā*), a sort of purification or consecration involving a transformation of the aspirant's personality, is regarded as a complement to, or even a substitute for, the previous initiation ceremony (the *upanayana* that all twice-born Hindus undergo at adolescence), which it strikingly resembles. Such religious groups integrate ancient, widespread ideas and customs of initiation into the framework of either the Vaishnava or Śaiva patterns of Hinduism. Vaishnavism emphasizes their character as an introduction to a life of devotion and as an entrance into closer contact with God, although happiness, knowledge, a long life, and a prospect of freedom from karma are also among the ideals to which they aspire. Śaivas are convinced of the absolute necessity of initiation for anyone desiring final liberation and require an initiation in accordance with their rituals. All communities agree that the authority to initiate belongs only to a qualified spiritual guide (guru), usually a Brahman, who has previously received the special *guru-dīkṣā* (initiation as a teacher) and is often regarded as representing God himself. The postulant is sometimes committed to a probationary period, to training in yoga mysticism, or to instruction in the esoteric meaning of the scriptures. The initiate receives a devotional name and is given the distinctive mantras of the community, which, because they are sacred, must never be misused.

Dīkṣā

There are many complicated forms of initiation: the Vaishnavas differentiate between the members of the four classes; the Śaivas and Tantrists take into account the natural aptitude and competency of the recipients and distinguish between first-grade initiates, who obtain access to God, and higher-grade initiates, who remain in a state of holiness.

Yoga. The initiate guided by his guru may apply himself to yoga (a "methodic exertion" of body and mind) in order to attain, through mortification, concentration, and meditation, a higher state of consciousness in which he may find the supreme knowledge, achieve spiritual autonomy, and realize his oneness with the Highest (or however the ultimate goal is conceived). Yoga may be atheistic or combined with various philosophical or religious currents. Every denomination attempted to implement yogic practices on a theoretical basis derived from its own teachings. There are many different forms of yoga, and the practices vary according to the stage of advancement of the adepts. All serious yogis, however, agree in disapproving the use of yogic methods for worldly purposes.

Sectarian symbols. The typical Hindu ascetic (sadhu) usually wears a distinctive mark (*puṇḍra*) on his forehead and often carries some symbol of his religion.

If he is a Vaishnava he might possess a discus (chakra) and a conch shell (*saṅkha*), replicas of Vishnu's flaming weapon and his instrument of beneficent power and

Symbols of Vishnu and Śiva

omnipresent protection, or a *śālagrāma* stone or a tulasi plant, which represent, respectively, Vishnu's essence and that of his spouse Lakṣmī. If he is a Śaiva, he might impersonate Śiva and carry a trident (*triśūla*), denoting empire and the irresistible force of transcendental reality; wear a small lingam; carry a human skull, showing that he is beyond the terror inspired by the transitoriness of the world; or smear his body with apotropaic (supposed to avert evil) and consecratory ashes. These emblems are sacred objects of worship because the divine presence, when invoked by mantras, is felt to be in them.

The attitude toward asceticism has always been ambivalent. On the one hand, there is a genuine regard for hermits and wandering ascetics and a desire to gain spiritual merit by feeding religious mendicants. On the other hand, the fact that fringe members of society may find a sort of respectable status among Śaiva ascetics often led to a decline in the moral reputation of the latter. (B.K.S./Ed.)

Cultural expressions: visual arts, theatre, and dance

The structure of Indian temples, the outward form of images, and indeed the very character of Indian art are largely determined by religion and a traditional view of the world, which penetrated the other provinces of culture and welded them into a homogeneous whole.

Indian art is highly symbolic. The much-developed ritual-religious symbolism presupposes the existence of a spiritual reality that, being in constant touch with phenomenal reality, may make its presence and influence felt and can also be approached through the symbols that belong to both spheres.

Mental and external fashioning of images

The production of objects of symbolic value is therefore more than a technique. The artisan must model a cult image after the ideal prototype that appears in his mind (in certain canonical forms) only when he has brought himself to a state of supranormal consciousness. After undergoing a process of spiritual transformation himself, he also transforms the material of which the image is to be made into a receptacle of divine power. Like the artisan, the worshiper (*sādhaka,* "the one who wishes to attain the goal") must grasp the esoteric meaning of a statue, picture, or pot and identify his or her self with the power residing in it. The usual offering, a handful of flowers, is the vehicle used to convey the worshiper's "life-breath" into the external image, which has already been transformed into an adequate internal vision of the same divine power.

TYPES OF SYMBOLS

If they know how to handle the symbols, the worshipers—who must achieve their object themselves and cannot come into contact with God unless they insistently invoke him—have at their disposition an instrument for utilizing the possibilities lying in the depths of their own subconscious as well as a key to the mysteries of the forces dominating the world.

Yantra and mandala. The general term for an "instrument [for controlling]" is yantra, which, while denoting in a wider sense cult images, pictures, and other such aids to worship, is often especially applied to ritual diagrams. Any yantra represents some aspect of the divine and enables devotees to worship it immediately within their hearts while identifying themselves with it. Except in its greater linear complication, a mandala does not differ from a yantra and both are drawn during a highly complex ritual in a purified and ritually consecrated place. The meaning and the use of both are similar, and they may be permanent or provisional. A mandala, delineating a consecrated place and protecting it against disintegrating forces represented in demoniac cycles, is the geometric projection of the universe, spatially and temporally reduced to its essential plan. It represents in a schematic form the whole drama of disintegration and reintegration, and the adept can use it to identify with the forces governing these. As in temple ritual, a vase is employed to receive the divine power so that it can be projected into the drawing and then into the person of the adept. Thus the mandala becomes a support for meditation, an instrument to provoke visions of the unseen. A good example of a mandala is the *śrīcakra,* "the Wheel of Śrī" (*i.e.,* of God's shakti) composed of four isosceles triangles with the apices upward, symbolizing Śiva, and five isosceles triangles with the apices downward, symbolizing Śakti; the nine triangles are of various sizes and intersect with one another. In the middle is the power point (*bindu*), visualizing the highest, the invisible, elusive centre from which the entire figure and the cosmos expand. The triangles are enclosed by two rows of (eight and 16) petals, representing the lotus of creation and reproductive vital force. The broken lines of the outer frame denote the figure to be a sanctuary with four openings to the regions of the universe. A "spiritual" foundation is provided by a yantra, called the mandala of the Puruṣa (spirit) of the site, that is also drawn on the site on which a temple is built. This rite is a reenactment of a variant of the myth of Puruṣa, an immortal primeval being who obstructed both worlds until he was subdued by the gods; the parts of his body became the spirits of the site.

Śrīcakra mandala

Lingam and yoni. One of the most common objects of worship, whether in temples or in the household cult, is the lingam (phallus). Often much stylized and an austere rather than literally sexual symbol, erect and representing the cosmic pillar, it emanates its all-producing energy to the four quarters of the universe. As the symbol of male creative energy it is frequently combined with its female counterpart (yoni), the latter forming the base from which the lingam rises. Although the lingam originally may have had no relation to Śiva, it has from ancient times been regarded as symbolizing Śiva's creative energy and is widely worshiped as his fundamental form.

Visual theology in icons. The beauty of cult objects contributes to their force as sacred instruments: their ornamentation facilitates the process of inviting the divine power into them. Statues of gods are not intended to imitate ideal human forms but to express the supernatural. A divine figure is a "likeness" (*pratimā*), a temporary benevolent or terrifying expression of some aspect of a god's nature. Iconographic handbooks attach great importance to the ideology behind images and reveal, for example, that Vishnu's eight arms stand for the four cardinal and intermediate points of the compass and that his four faces, illustrating the concept of God's fourfoldness, typify his strength, knowledge, lordship, and potency. The emblems express the qualities of their bearers—*e.g.,* a deadly weapon symbolizes destructive force, many-headedness omniscience. Much use is made of gestures (mudras), conventional devices for denoting activities that express an idea; thus, the raised right hand, in the "fear-not" gesture (*abhaya-mudrā*), bestows protection. Every iconographic detail has its own symbolic value, helping devotees to direct their energy to a deeper understanding of the various aspects of the divine and to proceed from external to internal worship. For many Indians, an installed and consecrated image becomes a container of concentrated divine energy; according to Hindu theists, it is an instrument for ennobling the worshiper who realizes God's presence in it.

Purposes of Hindu iconography

THE ARTS

Religious principles in sculpture and painting. The dance executed by Śiva as king of dancers (Naṭarāja), the visible symbol of the rhythm of the universe, represents God's five activities: he unfolds the universe out of the drum held in one of his right hands; he preserves it by uplifting his other right hand in *abhaya-mudrā;* he reabsorbs it with his upper left hand, which bears a tongue of flame; his transcendental essence is hidden behind the garb of apparitions, and grace is bestowed and release made visible by the foot that is held aloft and to which the hands are made to point; and the other foot, planted on the ground, gives an abode to the tired souls struggling in samsara. Another dance pose adopted by Śiva is the doomsday tandava, executed in his destructive Bhairava manifestation, usually with 10 arms and accompanied by Devī and demons. The related myth is that Śiva conquered a mighty elephant demon whom he forced to dance until he fell dead; then, wrapped in the blood-dripping skin of his victim, the god executed a horrendous dance of victory.

Images sustain the presence of the god: when Devī is

shown advancing against the buffalo demon, seated on her lion, she represents the affirmative forces of the universe and the triumph of divine power over wickedness. Male and female figures in uninterrupted embrace, as in Śaiva iconography, signify the union of opposites and the eternal process of generation. Lovers sculpted on temples are auspicious symbols on a par with foliage, water jars, and other representatives of fertility.

Like literature and the performing arts, the visual arts also contributed to the perpetuation of myths. Hindu sculpture tends to be less narrative than Buddhist, which delights in scenes from the Buddha's lives. In Hindu sculpture the tendency is toward hieratic poses of a god in a particular conventional stance (*mūrti*), which, once fixed, perpetuates itself. An icon is a frozen incident of a myth. For example, one *mūrti* (image) of Śiva is the "destruction of the elephant," in which Śiva appears dancing before and below a bloody elephant skin that he holds up with eight arms before the image of his horrified consort; the stance is the summary and reminder of his triumph over the elephant demon. A god may also appear in a characteristic pose while holding in his multitudinous hands his various emblems, on each of which hangs a story. Carvings, such as those that appear on temple chariots, tend to be more narrative; even more so are the miniature paintings of the Middle Ages. A favourite theme in the latter is the myth of the cowherd god Krishna and his love of the cowherd wives (*gopīs*), narrated in a series of paintings.

Religious organization of sacred architecture. Temples must be erected on a site that is *śubha* (*i.e.,* suitable, beautiful, auspicious, and near water) because the gods will not come to other places. However, temples are not necessarily designed to be congenial to their surroundings, because a manifestation of the sacred is an irruption, a break in phenomenal continuity. Temples are said to constitute an opening in the upward direction to ensure communication with the gods; they are visible representations of a cosmic pillar and their site is said to be a navel of the world. Their outward appearance must raise the expectation of meeting with God. Their erection is a reconstruction and reintegration of Puruṣa-Prajāpati, enabling him to continue his creative activity, and the finished monuments are symbols of the universe that is the unfolded One. The owner of the temple (*i.e.,* the individual or community that paid for its construction)—also called the sacrificer—participates in the process of reintegration and experiences his spiritual rebirth in the small cella, aptly called the "womb room" (*garbhagṛha*), by means of meditative contact with God's presence, symbolized or actualized in his consecrated image. The cella is in the centre of the temple above the navel—*i.e.,* the foundation stone; it may contain a jar filled with the creative power (shakti) that is identified with the goddess Earth (who bears and protects the monument), three lotus flowers, and three tortoises (of stone, silver, and gold) that represent Earth, atmosphere, and heaven. The tortoise is a manifestation of Vishnu bearing the cosmic pillar; the lotus is the symbol of the expansion of generative possibilities. The vertical axis or tube (coinciding with the cosmic pillar), which connects all parts of the building and is continued in the finial on the top, corresponds with the mystical vertical vein in the body of the worshiper through which his soul rises to unite itself with the Highest.

Requirements of temple architecture and planning

By courtesy of the Gujarat Museum Society, Ahmadabad, India



Krishna dancing with the *gopīs*, painting from western Rājasthān, *c.* 1610. In the N.C. Mehta Collection of the Gujarāt Museum Society, Ahmadābād, India.

The designing of Hindu temples, like that of religious images, was codified in the *Śilpa-śāstras* (craft textbooks), and every aspect of the design was believed to be symbolic of some feature of the cosmos. The idea of microcosmic symbolism is strong in Hinduism and comes from Vedic times; the Brāhmaṇa texts are replete with similar cosmic interpretations of the many features of the sacrifice. This same Vedic idea of the correspondence (*bandhu*) between microcosm and macrocosm was applied to the medieval temple, which was laid out geometrically to mirror the structure of the universe, with its four geometric quarters and a celestial roof. The temple also represents the mountain at the navel of the world and often somewhat resembles a mountain. On the periphery were carved the most worldly and diverse scenes, including luxurious celebrations of human life: battle scenes, hunts, circuses, animals, birds, as well as images of the gods.

Temple design as cosmic symbolism

The erotic scenes carved at Khajurāho in Madhya Pradesh and Konārak in Orissa express a general exuberance that may be an offering of thanksgiving to the gods who created all. However, that same swarming luxuriance of life in all of its aspects may also reflect the concern that one must set aside worldly temptations upon the threshhold of the sacred space of the temple, for the carvings only decorate the outside of the temple; at the centre, the sanctum sanctorum, there is little if any ornamentation, except for a stark symbol of the god or goddess. Thus, these carvings simultaneously express a celebration of samsara and a movement toward moksha.

Theatre and dance. Theatrical performances are also events that can be used to secure blessings and happiness; the element of recreation is indissolubly blended with edification and spiritual elevation. The structure and character of the classical Indian drama reveal its origin and function: it developed from the last part of a magico-religious ceremony, which survives as a ritual introduction, and begins and closes with benedictions. Drama is produced for festive occasions with a view to spiritual and religious success (*siddhi*), which must also be prompted by appropriate behaviour from the spectators; there must be a happy ending; the themes are borrowed from epic and legendary history; the development and unraveling of the plot are retarded; and the envy of malign influences is averted by the almost obligatory buffoon (*vidūṣaka,* "the spoiler"). There are also, in addition to films, which often use the same religious and mythic themes, *yātrās*, a combination of stage play and various festivities that have contributed much to the spread of the Purāṇic view of life.

Yātrās

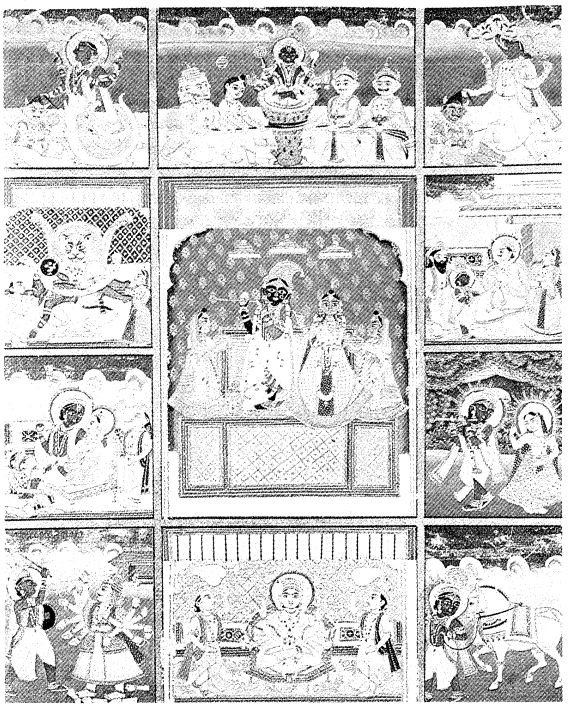
Dancing is not only an aesthetic pursuit but also a divine service. Hence there are halls for sacred dances annexed to some temples. The rhythmic movement has a compelling force, generating and concentrating power or releasing superfluous energy. It induces the experience of the divine and transforms the dancer into whatever he or she impersonates. Thus, many tribal dances consist of symbolic enactments of events (harvest, battles) in the hope that they will be accomplished successfully. Musicians and dancing girls accompany processions to expel the demons of cholera or cattle plague. Even today, religious themes and the various relations between humans and God are danced and made visual by the codified symbolic meanings of gestures and movements (see SOUTH ASIAN ARTS).

(A.L.B./J.A.v.B./W.Do./Ed.)

The place of Hinduism in world religions

HINDUISM AND OTHER RELIGIONS OF INDIAN ORIGIN

Hinduism, Buddhism, and Jainism originated out of the same milieu: the circles of world-renouncers of the 6th century BC. Although all share certain non-Vedic practices (such as renunciation itself and various yogic meditational techniques) and doctrines (such as the belief in rebirth and the goal of liberation from perpetual transmigration), they differ in the respect they accord to the Vedic tradition.



Vishnu with his 10 avatars (incarnations): Fish, Tortoise, Boar, Man-Lion, Dwarf, Rāma-with-the-Ax, King Rāma, Krishna, Buddha, and Kalkin. Painting from Jaipur, India, 19th century; in the Victoria and Albert Museum, London.

By courtesy of the Victoria and Albert Museum, London

Virtually all Hindus affirm the sacredness and authority of the Veda; Buddhists and Jains do not and therefore are regarded as less than orthodox by Hindus.

Buddhism. Although Buddhism did not interfere with Hindu customs and usages, allowing its adherents to approach Hindu or local supernatural powers for immediate favours, Hindu criticism of Buddhism came mainly from Brahman philosophers who opposed its adherents because they rejected the authority of the Veda and the Brahmans and the doctrine of the atman (soul) and because they admitted persons of any age and caste to monastic life. The spread of Buddhism was often regarded as an indication of degeneration. In the course of time, the Buddha was recognized as an incarnation of Vishnu, but this was often qualified by the addition that Vishnu assumed this form to mislead and destroy the enemies of the Veda, and this avatar is rarely worshiped. Buddhist emblems also were often ascribed to Vishnu or Śiva. Some Buddhist shrines have remained partly under the supervision of Hindu ascetics and are visited by pilgrims notwithstanding their much neglected condition.

After the rise of Buddhological studies in the West and the archaeological discoveries and restorations beginning at the end of the 19th century had made Indians more aware of the Indian origin of Buddhism, the Republic of India adopted the Buddhist emperor Aśoka's lion capital, marking the place of Buddha's first teaching, as its national emblem. The Buddha jubilee in 1956 was an occasion for enthusiastic celebrations. The number of Indian Buddhists has again increased, due mainly to the conversion of persons of low social rank who hope for higher social status as Buddhists than they were afforded as Hindus.

Jainism. With Jainism, which always remained an Indian religion, Hinduism has so much in common, especially in social institutions and ritual life, that nowadays Hindus tend to consider it a Hindu sect. Many Jains also are inclined to fraternization. The points of difference—*e.g.*, a stricter ahimsa practice and the absence of sacrifices for the deceased in Jainism—do not give offense to orthodox Hindus (see BUDDHISM, THE BUDDHA AND; JAINISM).

HINDUISM AND ISLĀM

Because Islām was so different from Hinduism in creed and institutions, it was neither absorbed nor powerful enough to make India a Muslim country. The religious situation created by the presence of its numerous adherents always had explosive potentialities: Muslims do not respect bovine life and regard Hindu cult practices as objectionable idolatry. Although Indian Muslims, with few exceptions, are of native descent, they are theoretically outcastes with whom dealings must remain restricted by formal rules; however, as with Christians, they are less polluting than the Hindu lower castes. The Islāmic way of life meets with opposition, and orthodox Muslims and Hindus do not ordinarily intermarry or dine together. This situation has had acute and even devastating consequences, but it does vary somewhat from region to region, from village to village, and from class to class. Very often mutual differences are accepted. Although they repudiate caste, Muslims often observe it in practice, and some have even retained their original caste organization after their conversion to Islām.

Syncretism of Hindu and Islāmic religion

Throughout centuries of close proximity and daily interaction, Hindus and Muslims have made efforts to accommodate the existence of the other religion within their own. One manifestation of such syncretism occurred among mystically inclined groups who believed that the one God, or "the universal principle," was the same regardless of whether it was called Allah or *brahman.* Various syntheses between the two religions, including Sikhism and other movements that emphasize nonsectarianism, have arisen in North India.

Those who, like Gandhi, could not understand the intolerance of orthodox Islām sympathized with the moderation and eclecticism of such groups. Most of the educated class, however, have always remained aware of the cleavage. To the Muslims—who, as part of an ecumenical community stretching over large parts of Asia and Africa, are concerned about the political and religious crisis of Islām since the late 19th century—the collapse of the Mughals after the Indian Mutiny (1857–58) was a severe blow that worsened relations with Hindus. This is particularly true because anti-Muslim tendencies had won ground since the renascent Hinduism of the Marāṭhā movement and in later times in the Arya Samaj (see above), while Muslims became self-assertive and even more determined to maintain their distinctive position. After the partition of the subcontinent into India and Pakistan—partly based on religious differences—and independence (1947), the political controversies between India and Pakistan constituted a further complication for relations between the religions.

HINDUISM AND CHRISTIANITY

The relations between Hinduism and Christianity have been shaped by unequal balances of political power and cultural influence. Although small communities of Christians have lived in South India since the middle of the 1st millennium, Christianity was widely introduced into the Indic subcontinent only in modern times by missionaries working under the auspices of British colonialism and imperialism. Their denigration of Hindu beliefs and practices—such as image worship and widow burning—provoked a Hindu response. Beginning in the 19th century and continuing to the present, a movement that might be called neo-Vedānta has emphasized the monism of certain Upanishads, decried "popular" Hindu "degenerations" such as the worship of idols, and acted as an agent of social reform, modernization, and dialogue between other world religions.

The relations between Hindus and Christians, then, have been complicated. Many Hindus are ready to accept the ethical teachings of the Gospels, particularly the Sermon on the Mount (whose influence on Gandhi is well-known) but reject the theological superstructure. Many adherents of bhakti movements—the Christian influence on which has been grossly exaggerated—feel that the Christian conceptions, which are regarded as a kind of bhakti, do not realize in God the multiplicity of human relations of love and service. Educated Hindus, though assimilating some Christian ideas, often regard missionary propaganda as an attack on their national genius and time-honoured institutions and take offense at what they regard as the disrespectful utterances of Christian missionary literature.

They are averse to the organization, the reliance on authorities, and the exclusiveness of Islām and Christianity, considering these as obstacles to harmonious cooperation. They subscribe to Gandhi's opinion that missionaries should confine their activities to humanitarian service. Since independence, conversion has indeed been viewed with disfavour by many influential Indians, who often also find in Hinduism what might be attractive in Christianity.

The Arya Samaj movement

Movements that advocate a Hindu theism designed to rival Islām and Christianity, like the Arya Samaj, make serious efforts to reconvert Christians to the Hindu community. People tolerate the proximity of Christian converts, even if they transgress Hindu taboos, provided they form a more or less separate community. Thus Christians often form castes or endogamous bodies analogous to castes. They sometimes are even admitted to temples to which untouchable Hindus have no entrance. In Malabar, due to their high economic position, Christians came to be practically equal with Brahmans. Nationalism has challenged the more serious-minded Indian Christians to express the genius of their faith in Indian modes and patterns. This has led, since 1921, to the emergence of Christian ashrams in the south. The dialogue between Hinduism and Christianity is more or less institutionalized at Bangalore in Karnātaka state, where the Christian Institute for the Study of Religion and Society is located. Its bulletin offers an opportunity for discussion between, for example, Christians and supporters of the Ramakrishna Mission.

(B.K.S./Ed.)

BIBLIOGRAPHY. Among the many overviews of Hinduism are THOMAS J. HOPKINS, *The Hindu Religious Tradition* (1971); DAVID R. KINSLEY, *Hinduism: A Cultural Perspective* (1982); R.C. ZAEHNER, *Hinduism,* (1962, reissued 1977); and LOUIS RENOU, *Religions of Ancient India* (1953, reissued 1972). An excellent survey of all aspects of pre-Muslim ancient India is A.L. BASHAM, *The Wonder That Was India,* 3rd rev. ed. (1967, reprinted 1985). More detailed and technical accounts may be found in JAN GONDA, *Die Religionen Indiens,* 2 vol. (1960–63); and in R.C. MAJUMDAR (ed.), *The History and Culture of the Indian People,* 11 vol. (1951–69). A basic resource for the study of Hinduism is the series "The History of Indian Literature," including the volumes by JAN GONDA, *Vedic Literature (Saṃhitās and Brāhmaṇas)* (1975), and *Medieval Religious Literature in Sanskrit* (1977); and by TEUN GOUDRIAAN and SANJUKTA GUPTA, *Hindu Tantric and Śākta Literature* (1981). A detailed outline is also provided by J.N. FARQUHAR, *An Outline of the Religious Literature of India* (1920, reprinted 1967). For historical overviews, consult VINCENT A. SMITH, *The Oxford History of India,* 4th ed. (1981); ROMILA THAPAR and PERCIVAL SPEAR, *A History of India,* 2 vol. (1965–66); and D.D. KOSAMBI, *The Culture and Civilisation of Ancient India in Historical Outline* (1965, reissued 1970; U.S. title, *Ancient India: A History of Its Culture and Civilization,* 1966, reissued 1969).

For the original sources of the principal texts of Hinduism in English translation, convenient collections include AINSLIE T. EMBREE (ed.), *The Hindu Tradition* (1966, reissued 1972); WENDY DONIGER O'FLAHERTY (ed. and trans.), *Textual Sources for the Study of Hinduism* (1988); AINSLIE T. EMBREE and STEPHEN N. HAY (eds.), *Sources of Indian Tradition,* 2nd ed., 2 vol. (1988); and R.C. ZAEHNER (ed. and trans.), *Hindu Scriptures* (1966). Compendiums of Hindu mythology in translation include WENDY DONIGER O'FLAHERTY (ed. and trans), *Hindu Myths: A Sourcebook* (1975); and CORNELIA DIMMITT and J.A.B. VAN BUITENEN (eds. and trans.), *Classical Hindu Mythology: A Reader in the Sanskrit Purāṇas* (1978).

Some of the individual textual classics of Hinduism have been translated and published in the series "Sacred Books of the East": F. MAX MÜLLER and HERMANN OLDENBERG (trans.), *Vedic Hymns,* 2 vol. (1891–97, reprinted 1979), selections from the Rigveda; MAURICE BLOOMFIELD (trans.), *Hymns of the Atharva-Veda: Together with Extracts from the Ritual Books and the Commentaries* (1897, reissued 1973); JULIUS EGGELING (trans.), *The Śatapatha-Brāhmana, According to the Text of the Mâdhyandina School,* 5 vol. (1882–1900, reprinted 1978); G. BÜHLER (trans.), *The Laws of Manu* (1886, reprinted 1971), with extracts from seven commentaries; JULIUS JOLLY (trans.), *The Institutes of Vishnu* (1880, reprinted 1965), and *The Minor Law-Books* (1889, reprinted 1965); GEORGE THIBAUT (trans.), *The Vedānta-Sūtras, with the Commentary by Rāmānuja,* 3 vol. (1890–1904, reprinted 1977); and HERMANN OLDENBERG (trans.), *The Grihya-Sutras: Rules of Vedic Domestic Ceremonies,* 2 vol. (1886–92, reissued 1973). See also WENDY DONIGER O'FLAHERTY (ed. and trans.), *The Rig Veda: An Anthology* (1981), a collection of 108 hymns. ROBERT ERNEST HUME (trans.), *The Thirteen Principal Upanishads,* 2nd ed. rev. (1931, reissued 1983), remains the best translation. R.C. ZAEHNER, *The Bhagavad-Gītā: With a Commentary Based on the Original Sources* (1969, reprinted 1973), is one of the best translations of this text. Translations of the epics include PRATAP CHANDRA ROY (trans.), *The Mahabharata of Krishna-Dwaipayana Vyasa,* 12 vol. (1883–96, reprinted 1981–82); J.A.B. VAN BUITENEN (ed. and trans.), *The Mahābhārata* (1973–); HARI PRASAD SHASTRI (trans.), *The Ramayana,* 3rd ed., 3 vol. (1976); and ROBERT P. GOLDMAN (trans.), *The Rāmāyaṇa of Vālmīki: An Epic of Ancient India* (1984–), with 2 vol. published by 1988. Translations of several *Purāṇa*s are available in J.L. SHASTRI, *Ancient Indian Tradition & Mythology* (1970–), with 35 vol. published by 1987; and in the *Purāṇas,* ed. by ANAND SWARUP GUPTA (1968–).

Although many of the vast sources for the vernacular literatures have not been translated, there are some, including: W. DOUGLAS P. HILL (trans.), *The Holy Lake of the Acts of Rāma* (1952), a translation of the *Rāmcaritmanas* of Tulasī Dās; CH. VAUDEVILLE (trans.), *Kabir,* vol. 1 (1974); LINDA HESS and SHUKDEV SINGH (trans.), *The Bījak of Kabir* (1983, reissued 1986); KENNETH E. BRYANT, *Poems to the Child-God: Structures and Strategies in the Poetry of Sūrdās* (1978); JOHN STRATTON HAWLEY, *Sūr Dās: Poet, Singer, Saint* (1984); A.K. RAMANUJAN (trans.), *Speaking of Śiva* (1973), *Hymns for the Drowning: Poems for Viṣṇu* (1981), and *Poems of Love and War: From the Eight Anthologies and the Ten Long Poems of Classical Tamil* (1985); and DAVID DEAN SHULMAN, *Tamil Temple Myths: Sacrifice and Divine Marriage in the South Indian Saiva Tradition* (1980), and *The King and the Clown in South Indian Myth and Poetry* (1985).

A good introduction to tribal and Hindu folklore is provided by VERRIER ELWIN (trans.), *Tribal Myths of Orissa* (1954, reprinted 1980), *Myths of Middle India* (1949, reprinted 1977), and *Myths of the North-East Frontier of India* (1958, reissued 1968). See also BRENDA E.F. BECK *et al.* (eds.), *Folktales of India* (1987); and STUART H. BLACKBURN and A.K. RAMANUJAN (eds.), *Another Harmony: New Essays on the Folklore of India* (1986).

For the prehistoric period and the Indus Valley civilization, see WALTER A. FAIRSERVIS, JR., *The Roots of Ancient India,* 2nd ed. rev. (1975); STUART PIGGOTT, *Prehistoric India to 1000 B.C.* (1950, reissued 1962); MORTIMER WHEELER, *The Indus Civilization,* 3rd ed. (1968); and JOHN MARSHALL, *Mohenjo-Daro and the Indus Civilization,* 3 vol. (1931, reprinted 1973). The best works on the Vedic religion include the essays found in J.C. HEESTERMAN, *The Inner Conflict of Tradition: Essays in Indian Ritual* (1985); see also FRITS STAAL, C.V. SOMAYAJIPAD, and M. ITTI RAVI NAMBUDIRI, *Agni: The Vedic Ritual of the Fire Alta,* 2 vols. (1983). The classic work on the Vedic sacrifice is SYLVAIN LÉVI, *La Doctrine du sacrifice dans les Brâhmanas* (1898, reissued 1966). Still useful are ARTHUR BERRIEDALE KEITH, *The Religion and Philosophy of the Veda and Upanishads,* 2 vol. (1925, reprinted 1971); and LOUIS RENOU, *Vedic India,* trans. from French (1957, reissued 1971). Works on the relations between Vedic religion and later Hinduism are JAN GONDA, *Change and Continuity in Indian Religion* (1965); MADELEINE BIARDEAU and CHARLES MALAMOUD, *Sacrifice dans l'Inde ancienne* (1976); and BRIAN K. SMITH, *Reflections on Resemblance, Ritual, and Religion* (1989).

The literature and teachings on dharma are presented in PANDURANG VAMAN KANE, *History of Dharmaśāstra (Ancient and Mediaeval Religions and Civil Law in India),* 2nd ed., 5 vol. in 8 (1968–77), an indispensible work. A summary of dharma is found in ROBERT LINGAT, *The Classical Law of India* (1973; originally published in French, 1967). The best study on yoga is MIRCEA ELIADE, *Yoga: Immortality and Freedom,* 2nd ed. (1969; originally published in French, 1954); also consult JEAN VARENNE, *Yoga and the Hindu Tradition* (1976; originally published in French, 1973). The doctrine of karma and rebirth as it is presented in the texts is examined in the essays collected in WENDY DONIGER O'FLAHERTY (ed.), *Karma and Rebirth in Classical Indian Traditions* (1980). CHARLES F. KEYES and E. VALENTINE DANIEL (eds.), *Karma: An Anthropological Inquiry* (1983), is also useful. Of the many works on the theoretical underpinnings of the caste system, the most influential of recent years has been the magnum opus by LOUIS DUMONT, *Homo Hierarchicus: The Caste System and Its Implications,* rev. ed. (1980; originally published in French, 1966). Also noteworthy is VEENA DAS, *Structure and Cognition: Aspects of Hindu Caste and Ritual,* 2nd ed. (1982).

For a convenient summary of the Hindu practice and ideology of image worship, consult DIANA L. ECK, *Darśan: Seeing the Divine Image in India,* 2nd rev. and enl. ed. (1985); and the essays in JOANNE PUNZO WAGHORNE, NORMAN CUTLER, and VASUDHA NARAYANAN (eds.), *Gods of Flesh/Gods of Stone: The Embodiment of Divinity in India* (1985). The best work on Hindu temples is STELLA KRAMRISCH, *The Hindu Temple,* 2 vol. (1946, reprinted 1976); see also GEORGE MICHELL, *The*

Hindu Temple: An Introduction to Its Meaning and Forms (1977). The practice of pilgrimage in general, and specifically pilgrimage to the holy city of Vārānasi, is treated in DIANA L. ECK, *Banaras: City of Light* (1982).

For an overview of the sects worshiping Vishnu or one of his forms, see JAN GONDA, *Aspects of Early Viṣṇuism* (1954, reissued 1969); SUVIRA JAISWAL, *The Origin and Development of Vaiṣṇavism: Vaiṣṇavism from 200 B.C. to A.D. 500,* 2nd rev. and enl. ed. (1981); and MILTON SINGER (ed.), *Krishna: Myths, Rites, and Attitudes* (1966, reprinted 1981). For an excellent study on Krishna, see ALF HILTEBEITEL, *The Ritual of Battle: Krishna in the Mahābhārata* (1976). Other important works include JOHN STRATTON HAWLEY, *Krishna, the Butter Thief* (1983); and DAVID R. KINSLEY, *The Divine Player: A Study of Kṛṣṇa līlā* (1979). SUSHIL KUMAR DE, *Early History of the Vaiṣṇava Faith and Movement in Bengal, from Sanskrit and Bengali Sources,* 2nd ed. (1961), is a historical approach.

For the history of Śaivism, see C.V. NARAYANA AYYAR (SADANANDA), *Origin and Early History of Śaivism in South India* (1936, reprinted 1974); and V.S. PATHAK, *History of Śaiva Cults in Northern India, from Inscriptions 700 A.D. to 1200 A.D.* (1960, reissued 1980). The mythology of Śiva is discussed in WENDY DONIGER O'FLAHERTY, *Asceticism and Eroticism in the Mythology of Śiva* (1973, reprinted as *Śiva, the Erotic Ascetic,* 1981); and in STELLA KRAMRISCH, *The Presence of Śiva* (1981).

The best overviews of Tantrism are AGEHANANDA BHARATI, *The Tantric Tradition* (1965, reprinted 1977); and EDWARD C. DIMOCK, JR., *The Place of the Hidden Moon: Erotic Mysticism in the Vaiṣṇavasahajiyā Cult of Bengal* (1966). For the worship of the goddess in her many forms, consult the essays in JOHN STRATTON HAWLEY and DONNA MARIE WULFF (eds.), *The Divine Consort: Rādhā and the Goddesses of India* (1982, reprinted 1986). See also THOMAS B. COBURN, *Devī Māhātmya: The Crystalization of the Goddess Tradition* (1984); CHEEVER MACKENZIE BROWN, *God as Mother: A Feminine Theology in India: An Historical and Theological Study of the Brahmavaivarta Purāna* (1974); DAVID KINSLEY, *Hindu Goddesses: Visions of the Divine Feminine in the Hindu Religious Tradition* (1986); and WENDY DONIGER O'FLAHERTY, *Women, Androgynes, and Other Mythical Beasts* (1980, reprinted 1982). An interesting and accessible comparison of certain themes in the worship of Krishna and the goddess is DAVID KINSLEY, *The Sword and the Flute: Kālī and Kṛṣṇa, Dark Visions of the Terrible and the Sublime in Hindu Mythology* (1975, reprinted 1977).

Emphasizing the anthropology of "popular" Hinduism are the works by LAWRENCE A. BABB, *The Divine Hierarchy: Popular Hinduism in Central India* (1975); MCKIM MARRIOTT (ed.), *Village India: Studies in the Little Community* (1955, reprinted 1986); MILTON SINGER (ed.), *Traditional India: Structure and Change* (1958, reissued 1976), and *When a Great Tradition Modernizes: An Anthropological Approach to Indian Civilization* (1972, reprinted 1980); and C.G. DIEHL, *Instrument and Purpose* (1956). A classic case history of the process known as Sanskritization is M.N. SRINIVAS, *Religion and Society Among the Coorgs of South India* (1952, reissued 1978). For a psychoanalytic approach to Hinduism, consult G. MORRIS CARSTAIRS, *The Twice-Born: A Study of a Community of High-Caste Hindus* (1957, reissued 1967); and SUDHIR KAKAR, *The Inner World: A Psycho-Analytic Study of Childhood and Society in India,* 2nd ed. rev. and enl. (1981, reprinted 1982), and *Shamans, Mystics and Doctors: A Psychological Inquiry into India and Its Healing Traditions* (1982, reissued 1984).

The standard work on the philosophical and theological aspects of various Hindu traditions is SURENDRANATH DASGUPTA, *A History of Indian Philosophy,* 5 vol. (1922–55, reprinted 1975). A fine series of volumes on Indian philosophy is KARL H. POTTER (comp.), *Encyclopedia of Indian Philosophies* (1977–), with 4 vol. published by 1987. For an analysis of one of these traditions, see JOHN BRAISTED CARMAN, *The Theology of Rāmānuja: An Essay in Interreligious Understanding* (1974).

Developments in the Hindu tradition as it confronted Western religions and modernity are covered in D.S. SARMA, *Studies in the Renaissance of Hinduism in the Nineteenth and Twentieth Centuries* (1944). Of special interest are the texts collected and translated by RICHARD FOX YOUNG, *Resistant Hinduism: Sanskrit Sources on Anti-Christian Apologetics in Early Nineteenth-Century India* (1981).

(W.Do./B.K.S.)

The Study of History

Modern historians aim to reconstruct a record of human activities and to achieve a more profound understanding of them. This conception of their task is quite recent, dating from the development in the late 18th and early 19th centuries of scientific history, cultivated largely by professional historians. It springs from an outlook that is very new in human experience: the assumption that the study of history is a natural, inevitable human activity. Before the late 18th century, historiography (the writing of history) did not stand at the centre of any civilization. History was almost never an important part of regular education, and it never claimed to provide an interpretation of human life as a whole. This was more appropriately the function of religion, of philosophy, even perhaps of poetry and other imaginative literature.

The article is divided into the following sections:

History of historiography 559
  Ancient historiography 559
    Greco-Roman era
    Early Christian era
    Early China
  Medieval historiography 562
    Europe from the 5th to the 11th century
    Europe from the 12th to the 14th century
  Byzantine historiography 564
  Muslim historiography 565
  Historiography in the European Renaissance 566
  Early modern historiography 567
  Historiography in the age of the Enlightenment 571
  Historiography in the 19th and 20th centuries 572
Methodology of historiography 573
  Source material 574
  Using source material 574
Ancillary fields 574
  Archaeology 574
    History of archaeology
    Fieldwork
    Interpretation
  Bibliography 579
    Descriptive bibliography
    Critical bibliography
  Chronology 582
    Chinese
    Japanese
    Indian
    Egyptian
    Babylonian and Assyrian
    Jewish
    Greek
    Roman
    Christian
    Muslim
    Pre-Columbian American
  Diplomatics 591
    History of the study of documents
    Diplomatic method
    Development and characteristics of chanceries
  Epigraphy 597
    Materials and techniques
    Inscriptions as historical source material
    Inscriptions as social and cultural records
    The use of inscriptions
    History of epigraphy
  Genealogy 606
    History of genealogical study
    Modern genealogy
  Paleography 609
    Types of writing materials
    Analysis of texts
  Sigillography 611
    Seals in antiquity
    Medieval European seals
    Modern use of seals
    Chinese and Japanese seals
  Textual criticism 614
    The materials of the investigation
    Critical methods
    History of textual criticism
Bibliography 620

History of historiography

ANCIENT HISTORIOGRAPHY

Greco-Roman era. The older, pre-18th-century outlook has been particularly well studied in the historiography of the ancient Greeks and Romans. But, although two of the most important ancient historians, Herodotus and Thucydides, wrote as early as the 5th century BC, when recorded Greek historiography was only just beginning, they had few successors of comparable quality. It is a symptom of the relative lack of importance attached in antiquity to this type of activity.

Ancient history was a branch of literature. The most appreciated historians were the writers who, like Thucydides, were able to touch on universal human problems or who, like the Roman author Tacitus (died *c.* AD 120), wrote in a dramatic way about important events or who, at least, attracted readers by their excellent style and skill in composition. Many of the works that lacked some of these literary qualities failed to survive.

About 1,000 ancient Greeks wrote in antiquity on historical subjects, but most of these writers are mere names. Many of the losses appear to have occurred in antiquity itself. Even historians of first rank have fared badly. Only in a few cases have complete texts of all their writings survived. Of the voluminous history of Polybius (covering originally the period 220–144 BC) only about one-third survives. Nearly half of Livy's Roman history (originally covering the period 753–9 BC) is lost. The text that remains is reasonably good only through the efforts of a group of Roman aristocrats who, in about AD 500, were trying to salvage the chief glories of Roman literature. A considerable part of Tacitus is missing, and the surviving portions of his *Annals* and *Histories* (originally AD 14–96) derive from two unique manuscripts.

Herodotus, whom the Roman statesman Cicero called "the father of history," came from the western coast of Asia Minor. The writers who preceded him were mainly Ionians from the Greek settlements in the same area. The origin of Greek historiography lies in the Ionian thought of the 6th century. The Ionian philosophers were doing something unprecedented: they were assuming that the universe is an intelligible whole and that through rational inquiries men might discover the general principles that govern it. Hecateus of Miletus, the most important Ionian predecessor of Herodotus, was applying the same critical spirit to the largely mythical Greek traditions when he wrote, early in the 5th century, "the stories of the Greeks are numerous and in my opinion ridiculous." Herodotus was more of a traditionalist, but he introduced his work as an "inquiry" (*historia*).

Egyptian and Babylonian historiography

A glance at the older historiography of the Egyptians, the Babylonians, and the other peoples of the ancient Near East will heighten one's appreciation of the novelty of the task undertaken by Herodotus. The kings of Egypt, of Babylonia and Assyria, and of the Hittites and the Persians all sought to preserve their glorious deeds for posterity in monumental inscriptions. The more im-

portant rulers also accumulated large archives, including both ordinary administrative documents and records specially commemorating their achievements. Some 20,000 clay tablets remain from the collections written for Ashurbanipal of Assyria (668–627 BC). Both in Egypt and in Babylonia lists of kings were kept in the temples, and these were sometimes supplemented by brief annals recording the principal events, though the hatred felt by certain rulers for their predecessors led to periodic destructions of older material. The exceptional meagreness of the narrative sources for Babylonian history before 747 BC seems due to the obliteration of the older annals by Nabonassar of Babylonia (ruled 747–734). Apart from changes in literary style, there was surprisingly little development over a period of more than 1,000 years in all these types of commemorative records. The inscriptions and temple records were normally intended to perpetuate the glory of the gods in whose service these rulers had accomplished great deeds. The names and dates of dynasties and of particular rulers can be reconstructed fairly adequately with the aid of these sources, but one cannot expect much accurate information about particular events. Nor, with rare exceptions, were those who had access to this material interested in using it to write continuous histories.

Herodotus and his immediate Ionian predecessors shared a very novel outlook. Its distinctive features were a lively curiosity and a capacity to treat sources in a critical spirit. Boundless curiosity about people and their diverse customs is one of the most endearing traits of Herodotus. Like other Greeks from western Asia Minor, he was particularly stimulated by contacts with the great Persian Empire, which offered opportunities for reasonably secure travel. The resultant immense widening of historical perspective is illustrated by a story told by Herodotus about Hecateus. When the latter assured the Egyptian priests at Thebes that he could trace his descent through 16 generations, the Egyptians showed him evidence of the descent of their high priests through 345 generations. Herodotus was the first to link his geographic inquiries with true history. His descriptions of the barbarian world that confronted the Greeks provided an introduction to the epic of the successful Greek resistance to the Persians.

The types of history written by the ancient Greeks and Romans influenced profoundly all subsequent historiography down to the 18th century. In order to interpret sympathetically this classical historiography, it is necessary to bear in mind the literary conventions that governed this branch of literature. The ancient Greeks distinguished between history and biography. The origin of both forms can be traced back to at least the 5th century BC, and the differences between them were observed throughout antiquity. The writer of history was supposed to aim at giving a true story, but the biographer was entitled to treat historical personages in a manner that resembled legend. There existed, of course, some exceptions. The lives of the early Roman emperors written by Suetonius in the 2nd century AD, while conforming to the traditional, topical arrangement of biographies, constitute an unusually valuable historical source, especially for Augustus, whose correspondence is repeatedly quoted. Yet another distinction was drawn between history and the study of "antiquities," to use a term employed by Varro (116–27 BC), perhaps the greatest of all the ancient Roman scholars. This distinction was already implicit in Aristotle's contemptuous dismissal of history (in his *Poetics*) as a branch of literature dealing with the particular rather than with things of general significance. The histories he condemned provided chronological narratives of wars and political events. Aristotle and his disciples were engaged in several enterprises that they regarded as something quite different from history. For example, they embarked on the study of the constitutions of all the Greek states. Such work was to be based on systematic inquiries. The student of the "antiquities" tried to use a wider range of evidence than the sources normally consulted by the ancient historians, and he arranged his results systematically by topics.

Ancient biography

In antiquity a writer of history was usually preoccupied at least as much with style as with content. A generation before Aristotle, the rules of rhetoric, as they might be applied to history, were fully elaborated by Isocrates, a teacher of rhetoric at Athens. Cicero tried (especially in his *De oratore,* 55 BC) to familiarize the Romans with these Isocratean precepts. History was to be written in a clear but solemn style, akin to fine oratory. The historian was to introduce all manner of literary embellishments but was also to stress the moral lessons of his story. At its worst this type of historiography could lead to serious misrepresentations of the past. Among the Roman historians, Livy (died AD 17) was an important practitioner of this kind of writing, which was particularly well suited to the patriotic myths that he was trying to immortalize, of a Rome that owed its magnificent destiny to the unique virtues of its citizens and the perfection of its antique institutions. Some outstanding historians, such as Polybius (2nd century BC) and Caesar (died 44 BC), eschewed these rhetorical precepts, but in all the ancient writers an important element of literary artifice was always present. This is one of the reasons why they offend modern standards, which demand absolute accuracy in the presentation of evidence. One of the most striking contrasts is the reluctance of the ancient historians to quote documents. Tacitus might rely heavily on the archives of the Roman Senate, but he never mentions his documentary sources. An inscription discovered at Lyons, France, preserves a speech delivered by the emperor Claudius to the Senate in AD 48, and it is clear that Tacitus utilized another version of the same text. His skill in using it is matched by the freedom with which he adapts it to suit his purpose.

The greatest and the most original achievement of the best Greek historians lay in their clear grasp of the need to distinguish truth from fiction and their conscious preoccupation with the methods of achieving this. This is admirably conveyed in a famous passage of Thucydides.

Methods of Thucydides

> And with reference to the narrative of events, far from permitting myself to derive it from the first source that came to hand, I did not even trust my own impressions, but it rests partly on what I saw myself, partly on what others saw for me, the accuracy of the report being always tried by the most severe and detailed tests possible. My conclusions have cost me some labour from the want of coincidence between accounts of the same occurrences by different eye-witnesses, arising sometimes from deficient memory, sometimes from deficient impartiality.

His practice did not fully live up to this ideal, however. The greatest of his Greek successors, Polybius, is reasonably impartial, except in his treatment of some of the events in Greece. Among the Romans, the writing of history was chiefly the preserve of members of the senatorial class, who almost invariably had some personal axes to grind. But the correctness of the rules formulated by Thucydides was accepted, in principle, by most ancient historians.

Thucydides had deliberately restricted himself to the history of his own time, and many of the subsequent ancient historians did likewise. They could depend on their own experience or could question well-informed contemporaries. The surviving fragments of Livy relating to his own lifetime (64/59 BC–AD 17) are much more vivid and convincing than the earlier books of his history (surviving today only down to 167 BC). The tendency to prefer contemporary history was strengthened by the practical bent of many of these writers. Several ancient historians were men of action familiar with warfare and politics. Interested in history as a source of instruction for statesmen, they could write with authority only about wars and political transactions of their own time. Polybius, the exiled Achaean general and a great traveller, derides unpractical, sedentary historians such as Timaeus, who had been writing about the peoples of the western Mediterranean without stirring for 50 years from Athens.

The historians of antiquity were much less skillful in dealing with noncontemporary history, for which they relied on older historians. Where none was to be found, they felt lost, as Livy complains in the early portions of his Roman history. The modern recourse to non-narrative sources was alien to the habits of most ancient historians. They were usually incapable of doing this successfully, just as they were ill equipped to discuss critically the sources used by the older writers.

Herodotus chose for his theme the successful resistance of the Greeks against the Persians at the beginning of the 5th century BC. Thucydides wrote about the Peloponnesian War, in which virtually all the Greek states became involved in the last decades of that century. These were limited subjects of obvious importance for which it was possible to find ample evidence. The strength of the ancient historians lay precisely in imposing an interesting pattern on the events of a selected period, usually contemporary or fairly recent, for which they had manageable sources. The best of them could thereby achieve a sense of dramatic unity and produce literary masterpieces. The speeches that Thucydides invented for some of the main protagonists in his story are artistically the most satisfying parts of his work, and at times they even seem to recapture the spirit of what might have been said on these occasions. In a superb writer like Tacitus, whose political career had included long periods of frustration and insecurity, one does not look for impartiality or for scrupulous truthfulness but, rather, for fascinating insights into what the development of Roman imperial power from Augustus to Domitian (the period AD 14–96) meant to the proud, sophisticated Roman aristocracy for whom he was writing.

Classical study of "antiquities"

The study of "antiquities," as opposed to narrative history, did not normally produce works of literary merit, and this is probably the main reason why most of them disappeared. One important group of such writings originated with Aristotle and his collaborators, writing in the third quarter of the 4th century BC. They were interested in both literary "antiquities" and in the systematic study of the constitutions of Greek states. They had described 158 different constitutions, though only their account of Athens now survives. A comparison of its two main parts illustrates the contrast between the deficiencies of ancient historiography and the impressive achievements of the antiquarian researchers. In the introductory, historical section, Aristotle was baffled by the problem of dealing with the fairly remote past. For each particular period he tried to follow some contemporary sources. The resultant juxtaposition of several writers differing widely in their political outlook produced an account full of contradictions. The second part, however, containing a systematic description of the Athenian constitution, is a masterpiece of shrewd analysis, as are the empirical portions of Aristotle's *Politics* (Books IV–VI), which are based on a wealth of concrete examples derived from the different Greek states.

Aristotle inspired in the 3rd and 2nd centuries BC a great mass of philological and antiquarian research. The most important scholars were to be found in the new Hellenistic states, especially at Alexandria in Egypt and at Pergamum in Asia Minor. Among the surviving Hellenistic fragments, there are commentaries on Herodotus and Thucydides. The Hellenistic scholars were interested in many subjects connected with history and did pioneering work in chronology, geography, and topography. They were accustomed to using every kind of source and to quoting documents extensively. Their greatest Roman disciple was Varro, who tried to recover all the vestiges of the old Roman society and to make a systematic survey of Roman life based on the evidence provided by language, literature, religion, and ancient customs. Most of his writings have been lost, but he supplied the conjectural (though incorrect) date of 753 BC for the foundation of Rome and knowledge of the probable boundaries between some of the groups whose union produced the city of Rome. Unfortunately, antiquarian researches of such penetrating nature were almost never applied in antiquity to the writing of narrative histories.

Early Christian era. The triumph of Christianity in the Roman Empire during the 4th century assured the predominance of a type of historiography radically different from the works of the pagan Greek and Roman historians. Its origins were Jewish. The Jews were the only people of antiquity who had the supreme religious duty of remembering the past because their traditional histories commemorated the working out of God's plan for his chosen people. By contrast, no Greek ever heard his gods ordering him to remember. It was the duty of every Jew to be familiar with the Jewish sacred writings, which were ultimately gathered into what became the Old Testament. The writers of these biblical books only gave an authoritative version of what everybody was supposed to know, and they were only concerned with the selection of such facts as seemed relevant in interpreting God's purpose. In addition, the Jews also cherished unwritten traditions. To quote Josephus, a Jewish historian of the 1st century AD, "what had not been written down, was yet entrusted to the collective memory of the people of Israel and especially of its priests."

Purpose of Jewish histories

The Christians took over the Old Testament and added to it an additional body of sacred history. The writers of the four Gospels included in the New Testament were bearing witness to assured truths that the faithful ought to know, and no convincing reconstruction of historical facts is possible from these books of the New Testament. The only avowedly historical book in it is the Acts of the Apostles. The New Testament as a whole represents merely a selection from the early Christian writings. It includes only what conformed to the doctrine of the church when, later on, that doctrine became fixed in one form. Between the Acts of the Apostles, dating probably from the late 1st century, and the writings of Eusebius of Caesarea (died *c.* 340) and his contemporaries in the first quarter of the 4th century, there is an almost complete gap in Christian historiography.

For the Christian writers the story of Jesus, as recorded in the Gospels, represented the fulfillment of the prophecies that could be found in various parts of the Old Testament. The Jewish part of the Bible also assured for Christianity the authority of a long antiquity. The history contained in the two parts of the Bible, now indissolubly linked together, became the only authentic record of God's revelation for mankind, dwarfing into insignificance all the records of other peoples and religious groups. The concept of a universal history had not been wholly unknown to the pagan world, but the Christians were the first to apply it effectively. Christian history had to be a universal history, though of a very peculiar sort, where only one sequence of privileged events, Jewish and Christian, deserved detailed record. The Christian claims must have seemed more extravagant to the pagans than even the Jewish ones. Thus Eusebius stated that the Christians were, in fact, born with the world, anticipating St. Augustine's vision of the city of God existing since the beginning of time.

Universal chronologies

In defending their religion against hostile critics, the early Christians were forced to fit some pagan history into their universal scheme. This was achieved by means of universal chronologies from the creation of the world to each writer's own time. The events of Jewish and Christian history were thus synchronized with the main dates of the pagan myth and history. Sextus Julius Africanus, who wrote in the early 3rd century, is the first Christian writer known to have attempted this feat. He allotted 6,000 years to the whole span of human history and placed the birth of Christ in the year 5500 from the creation of the world. This work provided the model for the more elaborate *Chronographia* (*Chronicle*) of Eusebius. It became the foundation for a long succession of Greek chronographies produced by Byzantine writers. A Latin adaptation by St. Jerome (died 419/420) was immensely influential in western Europe for more than 1,000 years. A modern scholar is filled with mingled admiration and despair at the ingenuity of Eusebius and of his more eminent successors and at the absurdity of many of their conclusions. But they did originate and impose on the world a unified scheme of universal chronology. The dating from the birth of Christ was introduced by Dionysius Exiguus, who wrote at Rome in the early 6th century, and it was successfully popularized in the 8th century by the English historian Bede.

The writing of history of their own time was not an essential task for the Christians of the 4th and 5th centuries. When they did so, they wrote primarily in defense of their religion against the pagan world or against rival Christian groups branded as heretical. All these histories belong to religious apologetics. They suffer from inevitable distortions in the choice of what should be mentioned and what must be suppressed, and they are often excessively unfair to outsiders and opponents. These faults were not uncom-

mon among the classical historians, though the Christians were somewhat unusual in their extreme conviction that they alone must be right. A comparison between the Christian historians and an outstanding pagan writer, such as Ammianus Marcellinus (second half of the 4th century), who was very ready to admire those Christians who merited it, brings out the intolerance and narrowness of outlook of his Christian contemporaries.

Eusebius was the earliest and the most important of the Christian historians of the 4th century. He is quite frank about the practical and apologetic aims of his *Historia ecclesiastica* (written 312–324; *Ecclesiastical History*) designed to show how, through a long series of acts of Divine Providence, a Christian empire was finally brought into existence by Constantine. He admits that "we shall introduce into this history in general only those events which may be useful first to ourselves and afterward, to posterity." This work, like his other historical writings, is a mixture of devout fiction and invaluable detail. But there is plenty of the latter in *Ecclesiastical History*. Contrary to the usual practice of the ancient historians, Eusebius tries to specify his sources, and he quotes from them extensively in order to document as fully as possible the developments that resulted in the triumph of Christianity. He provided in this respect a valuable model for his medieval successors. The most astonishing thing about Eusebius was his capacity to handle his sources critically, in matters where it seemed permissible to do so. In one passage of his *Chronicle* he sets aside the authority of St. Paul in favour of a piece of evidence contained in the Book of Judges. In later patristic literature nothing similar is found.

Biography, as it was habitually written in antiquity, could be readily adapted to Christian purposes. St. Jerome modelled himself on Suetonius in compiling the lives of 135 Christian writers (written in 392) as a way of demonstrating the high level of culture attained by his coreligionists. The ancient biographers had freely mingled fact with fiction for the edification of their readers and could be readily imitated by the writers of the lives of Christian saints. The life of St. Anthony of Egypt by St. Athanasius (mid-4th century) set the pattern for this most popular type of medieval literature.

St. Augustine

St. Augustine, the greatest of the Latin Church Fathers of the 4th and 5th centuries, was certainly not concerned with writing of history in any ordinary sense of the term. In his *De civitate Dei* (*City of God*) he might invoke historical evidence to demonstrate the utter degradation of all the non-Christian societies, and he encouraged his pupil Orosius to develop this theme more fully in the latter's *Historiarum libri VII adversus paganos* (*Seven Books of History Against the Pagans,* to 417). Nearly 200 manuscripts of Orosius have survived, testifying to the immense popularity of his work in the Middle Ages. Augustine's greatest influence on historiography lay in his main message. His vision of the divine and the earthly cities confronting each other dominated the outlook of all the medieval Christian thinkers and profoundly affected their treatment of history. Within that divine plan for the world, purely secular history seemed an insignificant thing.

Early China. The preservation of some records of historical events can be traced in China to at least the early part of the 1st millennium BC. Confucius (551–479 BC) was credited, rightly or wrongly, in the later Chinese tradition with editing the annals of his native state of Lu. But

The 'Historical Records" of ancient China

the appearance of the first works fully deserving the name of histories resulted from the unification of China under a single ruler in 221 BC. The first such work to survive, the *Shih chi* ("Historical Records"), dates from *c.* 85 BC. Its author, Ssu-ma Ch'ien, is quite justifiably called the father of Chinese historiography. His history exhibits many of the main features of the later Chinese official histories as they continued to be written down to the deposition of the last Chinese imperial dynasty in 1911. Within this fairly unified tradition, China produced a mass of historical writings unequalled by any other country before modern times. Until the late 19th century, Japanese historiography formed an offshoot of this tradition.

Chinese scholars showed an interest in the history of China from the earliest times. According to the Chinese conception, history makes sense only if it can furnish practical directives for action or supply correct information upon which action can wisely be based. All the schools of Chinese thought quoted the lessons of history. Confucius, with his stress on the moral content of these lessons, formed part of this universal belief in the value of history. One of the duties inculcated by him was the scrupulous transmission of authentic records. When, some centuries after his death, the unified Imperial state began to recruit its bureaucracy among the Confucian scholars, the recording of all the necessary information and the careful preservation of records became one of the main functions of the Chinese government, both centrally and locally. A long series of official histories and of records connected with them has survived from the time of the T'ang dynasty (618–907) onward. From then on, the great bulk of Chinese history was written by bureaucrats for bureaucrats. From a practical point of view this immense body of historical writings fulfilled a very useful purpose. Such histories were bound to be highly stereotyped and restricted in content to what interested the higher officialdom. It is easy to condemn it by modern Western standards for its excessive preoccupation with concrete details and inability to produce works of wider synthesis. But this Chinese tradition did gradually evolve in the direction of greater rationality and subtlety. Its scope widened as the sphere of government expanded. Furthermore, within this tradition there appeared from time to time writers of genius, men of bold critical spirit, genuine historical insight, and overriding integrity. One of the greatest was Liu Chih-chi (661–721), the writer of the *Shih t'ung,* the first thorough treatise in Chinese, or any other language, on historical method, which also constituted in effect a history of Chinese historiography. He had a successor in Ssu-ma Kuang (1019–86), the author of the first fairly comprehensive general history of China (covering the years 403 BC–AD 959). In the 17th century a remarkable group of historical scholars virtually founded a school of critical Chinese philology. None of these writers succeeded in radically transforming Chinese historiography, but they created an increasingly sophisticated and critical tradition. Their successors in the 20th century assimilated some valuable features of modern Western historiography.

MEDIEVAL HISTORIOGRAPHY

Europe from the 5th to the 11th century. The period stretching from the 5th to the 11th century was a time of very profound cultural decline in regions that had once constituted the western half of the Roman Empire. Almost all the inhabitants of these provinces again became illiterate. There are long periods for which there are virtually no narrative sources, and the bulk of surviving historical writings consists merely of meagre factual annals. Virtually all the writers were ecclesiastics, in marked contrast to the Byzantine lands, where a strong tradition of lay historiography persisted throughout the Middle Ages. The annalists and chroniclers of the West were predominantly monks, and their lack of experience of the secular world outside their cloisters made them into blinkered and unpractical historians. This was true even of Bede, an Anglo-Saxon monk, who was by far the greatest historian of the early Middle Ages.

All the historians of this period were seriously affected by the cultural decline around them. They were having to write in part for a more uncultured audience. Sulpicius Severus, probably the best Western historian of the early 5th century, still intended his *Chronica* (to 403) for educated Roman Christians, but his life of St. Martin of Tours is a piece of medieval hagiography. This model could inspire lives full of folklore and miracle, from which the real human personalities of the saints were almost wholly absent. The same duality of purpose is a notable feature of Bede's voluminous writings. He explicitly recognized that he must adapt himself to his audience when he explained that he was writing in a simple Latin style so that he might be more easily understood by his Anglo-Saxon readers. There is a marked contrast of tone between his theological and his historical writings. As a theologian, Bede follows Eusebius and the earlier Church Fathers in not exaggerat-

ing the frequency of miracles and in believing that they were most common in the earliest days of Christianity. But Bede's lives of the English saints and his *Historia ecclesiastica gentis Anglorum* (*Ecclesiastical History of the English People*), covering chiefly the years 597–731, are full of miracles and visions. There is one or other on almost every page. It is possible that some of these incidents were included by Bede because he thought that his readers expected mentions of these familiar, traditional stories.

Bede's *Ecclesiastical History*

In preparing his historical works, Bede not only took great care to assemble the widest possible collection of sources but also tells the reader what he is using. In dedicating his *Ecclesiastical History* to King Ceolwulf of Northumbria, he requests that

> in order to remove all occasions of doubt about those things I have written, either in your mind or in the minds of any others who listen to or read this history, I will make it my business to state briefly from what sources I have gained my information.

An impressive list follows, including mentions of documents copied for him by friends at Rome, Canterbury, and other places. Like Eusebius, on whom Bede modelled himself, he quotes some of the documents integrally. Bede's methods of securing and recording information are so similar to the practices of modern historians and the judicious tone of his writing is so impressive that the reader is almost taken in into treating him as if he were a modern scholar. But Bede's *Ecclesiastical History* was written as a work of edification in order to strengthen the faith of his readers in Divine Providence, through which, as he saw it, his Anglo-Saxon countrymen had been converted to Christianity. All matters not connected with his main theme are ignored. Bede's handling of evidence on subjects that he regarded as embarrassing inspires mistrust. But these are small matters in comparison with the enormous mass of information that he alone has preserved and the encouragement that Bede continued to give for many centuries to the writing of history.

The influence of Bede and other Anglo-Saxon scholars was greatly felt during the later 8th and the 9th centuries in the Frankish kingdom, where under Charlemagne and his successor, Louis the Pious, there was a modest revival of historical writing. Besides the annals kept at various monasteries, which tended to convey information in a manner that suited the Frankish rulers, there were a few more ambitious ventures. The important *Historia Langobardorum* (*History of the Lombards*), written *c.* 774–785 by Paulus Diaconus, or Paul the Deacon, was the work of one of the best educated men of the time. Nithard, a grandson of Charlemagne, left an invaluable narrative of the disintegration of the Carolingian state during his lifetime. The work that exerted the greatest influence on the medieval writers of biographies was Einhard's *Vita Karoli Magni* (written *c.* 830–833; *Life of Charlemagne*). The author was a leading official and a close companion of Charles, and his work was naturally intended as a eulogy of the great king. Einhard says that Charlemagne retreated safely from Spain, returning with his army safe and sound, except that on a ridge of the Pyrenees, on the way home, he happened to experience some small effects of Gascon perfidy. Nobody would gather from this that the Franks had narrowly escaped a major disaster. Einhard was merely echoing the story told in the semiofficial contemporary annals. Another source of distortion was Einhard's use of a classical model, the *Lives of the Caesars* by Suetonius. The subject headings under which he described Charles and even the very words used were partly borrowed from the lives of Roman emperors, but his Charlemagne is probably in essentials an authentic and credible portrait.

If bulk alone is to be taken as a criterion, annals were the main product of medieval historiography. The annalist merely sets down the most important events of the current year. In the case of the earliest medieval annals, the events were often noted down in Easter tables, in the blank spaces between the dates calculated for the forthcoming Easters. Such paschal annals would be extremely brief. When, as often happened, annals came to be written down in separate manuscripts, distinct from the Easter tables, there was room for the expansion of individual entries. In either case, the resultant annals cannot be regarded as history since the events are necessarily recorded in isolation. But they preserve in a right order the essential facts, which could be rearranged into a continuous narrative. Such a narrative, if it still followed the chronological arrangement of its various annalistic sources, should properly be termed a chronicle.

Medieval historians show little awareness of the process of historical change. They were unable to imagine that any earlier age was substantially different from their own. The unawareness of the meaning of anachronism helps to explain the strange wanderings of medieval annals and chronicles. If a religious community wanted to acquire a historical narrative, it copied some work that happened to be most readily accessible. A continuation might then be added at the manuscript's new abode, and, later on, this composite version might be copied and further altered by a succession of other writers. Hence there are at least six main versions of the annals known as the *Anglo-Saxon Chronicle*. They all derive from the annals kept down to 892 at Winchester, the West Saxon capital. Thereafter, copies were acquired by religious centres in the most diverse parts of England, and one manuscript was being kept up to date at the abbey of Peterborough as late as 1154. An extreme case of wanderings is represented by the annals of the cathedral church of Cracow, the medieval Polish capital. The first section is based on Orosius, the next comprises annals beginning with the death of Bede and containing notices of Frankish and German events, while the Polish section starts with the conversion of Poland to Christianity (965–966) and ends in the 13th century.

The *Anglo-Saxon Chronicle*

Europe from the 12th to the 14th century. Historians are accustomed to regarding the late 11th and 12th centuries as an age of intensified progress in culture and learning; this development, however, did not greatly affect historiography. There was a modest revival of interest in some of the ancient Latin writers, but would-be historians were unsure which ancient models they ought to imitate. A whole series of attempts was made to apply to other races the theme in Virgil's *Aeneid* of a noble group of people guided by the gods toward a splendid destiny. The first essential step was to establish the descent of one's nation from the ancient Trojans and then to trace subsequent history through a series of heroic conquests. The most ambitious of these writings was the *Historia regum Britanniae* (*History of the Kings of Britain*), by Geoffrey of Monmouth (died 1155), which attempted to establish for the Celts a historical destiny greater than any other. Although some, even contemporary, readers were not deceived by the work, and William of Newburgh, one of the best English historians of the 12th century, denounced it as a tissue of absurdities, many seriously accepted it as history.

With a few exceptions, the ablest minds of the 12th century were attracted into enterprises that ignored history; they were more concerned with systematization of thought and with philosophical speculations. One of the exceptions was Otto, bishop of Freising, in Bavaria. He was a grandson of the Holy Roman emperor Henry IV. He received the best education that his age could give, but he was also briefly a Cistercian monk during the most austere period of that order's history. Otto was torn between conflicting impulses to seek the city of God as the only reality and yet to hope for the progress of the German empire. Out of this conflict came his first work, *Chronica* (*The Two Cities*), a chronicle of world history to 1146, perhaps the most profound medieval attempt at a Christian philosophy of history. As Otto himself confessed, it was composed "in bitterness of spirit . . . in the manner of tragedy." The election in 1152 of his nephew and friend Frederick Barbarossa, as emperor, filled Otto with a new elation. The excellence of his second work, *Gesta Friderici I imperatoris* (*The Deeds of Frederick Barbarossa*), derives in a considerable measure from a quality rare in medieval historians, a sense of optimistic belief in the value of writing history because it might become a record of human progress. *The Deeds of Frederick Barbarossa* contains a penetrating analysis of the problems encountered by the

The works of Otto of Freising

German rulers in trying to rule the precociously urbanized Italian society.

As in antiquity, the best medieval works were accounts of contemporary history by men who had participated in the events that they were describing. It is, however, very significant that some of the writers that are prized most highly today survive in only very few manuscripts and were presumably not appreciated by most of their contemporaries. One such work was the *Historia pontificalis* ("Pontifical History") covering the period 1148–52, of John of Salisbury, one of the most accomplished scholars of his age, who was writing about the period when he was in the papal service. Another instance of undeserved neglect is furnished by the *Liber de regno Siciliae* ("Book of the Kingdom of Sicily") covering the period 1154–69, written by an anonymous member of the Sicilian court.

Unlike the ancient historians, the medieval writers of contemporary history had no inhibitions about extensively quoting official documents. In England, a succession of writers preserved a large quantity of such texts. Roger of Hoveden was, in the last quarter of the 12th century, treated by the English kings as a kind of court historian. He preserved valuable legal and administrative records with which he was familiar through his activities as a royal official and justice. Matthew Paris, the most important English monastic historian of the 13th century, was highly regarded by King Henry III and had excellent sources of information. He left behind a collection of transcripts of royal and ecclesiastical documents that today fills a large printed volume. Some writers made their chronicles into an anthology of official records, thinly connected by the author's brief comments. Such is the chronicle of Robert of Avesbury, consisting mainly of the military dispatches of King Edward III and other interesting documents to 1356. Another variant of the same method was for a wholly mediocre chronicle to incorporate exciting pieces of eyewitness narratives by other writers. A dull English monastic product of the late 14th century, the *Anonimalle Chronicle,* includes a narrative of the Peasants' Revolt of 1381, which is one of the most dramatic and interesting eyewitness accounts to be found in medieval historiography.

The most popular histories of the 13th and 14th centuries were encyclopaedic compilations giving all the important facts neatly arranged under the dates of popes, emperors, and other rulers. There were even more ambitious ventures aiming at summarizing all the important facts from all the different branches of human activity. The Dominican Order, created at the beginning of the 13th century, was especially concerned with producing such aids for the dissemination of useful knowledge. The best known of these Dominican works is the immense *Speculum historiale* ("Mirror of History"), by Vincent of Beauvais, written under the patronage of King Louis IX of France. It is a compilation made up of excerpts from many authors.

The 13th and 14th centuries were not a period of any fundamental innovations in the techniques and nature of historiography, but there was a growing diversity of types of historical writing. Very detailed, chatty narratives multiplied, often badly organized and inaccurate, but conveying the authentic atmosphere of the times and vividly portraying leading personalities. Such were the St. Albans chronicles of Matthew Paris (to 1259), the reminiscences of Joinville about St. Louis during the Seventh Crusade (1248–54), the Lombard chronicle of Fra Salimbene (to 1287), or the vast history of the first part of the Hundred Years' War written in the second half of the 14th century by Froissart. Memoirs and histories written in vernacular languages, such as those of Joinville and Froissart, came to be quite common. Laymen began to write histories. Some were great men, like Geoffroi de Villehardouin, one of the leaders of the Fourth Crusade (which captured Constantinople 1202–04), of which he wrote an account. Important urban chronicles began to appear, such as the Florentine chronicle of Giovanni Villani, with its invaluable statistics of Florentine population and activities around 1338. The extraordinary personality of St. Francis, who died in 1226, inspired lives of him more convincingly human than any previous medieval biographies of saints.

Vernacular histories

The Humanist historians of the 15th century tried to make a deliberate break with the tradition of medieval historiography. By their insistence on a more coherent arrangement of subject matter, by their superior critical outlook, and, above all, by their much more accurate awareness of the process of historical change, they had introduced innovations of fundamental importance. In part they owed their grasp of these new possibilities to the influence of Byzantine scholars. In historiography, as in other matters, the new humanistic scholarship was a joint product of Western and Byzantine traditions.

BYZANTINE HISTORIOGRAPHY

During the millennium that elapsed between the collapse of the Roman Empire in the West in the 5th century AD and the Italian Renaissance of the 15th century, in no part of Europe did the writers of history consistently maintain as high a standard of achievement as in the Byzantine Empire. Parts of the 7th and the 8th centuries form lengthy gaps in the record of Byzantine historiography, but this seems mainly to be the result of subsequent losses of manuscripts. When, in the middle of the 9th century, Photius, future patriarch of Constantinople, compiled a record of some 280 books that he had read, he mentioned works of 33 Greek historians, dating mostly from the late Roman Empire and the Byzantine period, 20 of which are now lost. But, among the Byzantines of the 7th and 8th centuries, there was certainly no parallel to the Dark Ages in western Europe.

Sources of Byzantine historiography

The Byzantine historians were heirs to the combined traditions of classical Greek writing, of the subsequent Hellenistic historiography, and of the Christian historical writing of the 4th century. Few ancient Latin historians were ever translated into Greek, and their influence on the Byzantines was, therefore, very slight. The older classical Greek historians provided the Byzantines with their cherished models of language and style. Like all educated Byzantines, the historians continued for a millennium to write in a literary language that soon became unintelligible to the vast majority of their compatriots. Hence, from the 6th century onward, there appeared, side by side with the learned historiography, a succession of popular chronicles written in the ordinary language. Most of these popular writings form—in their prejudice, ignorance, and crudity—a startling contrast to the works of the more eminent classicizing historians, but they do provide valuable glimpses of the sort of hagiographical history, more religious myth than sober fact, that ordinary Byzantines apparently wanted to read.

Herodotus and Thucydides were frequently invoked by Byzantine historians as models of fine prose. The influence of these two writers on the substance of what was written usually remained slight and superficial, however. The only Byzantine writers who seriously modelled themselves on these two oldest Greek historians wrote during the 15th century. The earlier Byzantine historians owed most to Polybius and to the Greek biographer Plutarch (died *c.* AD 119), the two Hellenistic writers who had the greatest influence on Byzantine notions of how history and historical biography should be written.

Like Polybius, the majority of Byzantine historians, including most of the best ones, perferred to write about their own times; and within these limits they produced some real masterpieces. Unlike the majority of the ancient historians, Polybius had included much autobiographical detail, and his influence reinforced the readiness of the Byzantine historians to talk about themselves, thus providing abundant information about several of these authors. Their histories are likely to be one-sided and full of details about what interested them, while remaining silent about a great mass of other contemporary happenings. They are frequently gossipy and patently prejudiced, inspiring much less confidence than the austere, impartial writings of authors such as Thucydides. This is one of the main reasons why the Byzantine historians have often been excessively underestimated by modern readers. The bulk of the Byzantine contemporary histories were written by statesmen, high officials, and prelates—men with access

to important information. They have to be used critically and cautiously but can be immensely valuable.

Priscus of Panium (*c.* 450), a member of a Byzantine embassy to Attila's camp, is the best source of information about that terrible king of the Huns and his followers. A century later, the reconquest of Vandal Africa and of Ostrogothic Italy by the emperor Justinian was the main theme of the *History of the Wars* of Procopius, a leading civilian adviser of Belisarius, the Byzantine commander. Subsequently, Procopius also wrote a *Historia arcana* (*Secret History*), containing a horrible indictment of the activities of Justinian and Belisarius. Many of his details about the corruption at court and the oppressive nature of the government may be substantially correct. In the 11th century Michael Psellus, who wrote a history of his own times, was a leading Byzantine scholar and official, for a time even the chief adviser of emperors. His *Chronographia* is concerned almost entirely with the happenings at the Byzantine court and is one of the most gossipy and amusing narratives ever written on such a subject. His psychological insight and his lively and subtle style delighted the educated Byzantines. Anna Comnena, the daughter and biographer of the emperor Alexius I, greatly admired Psellus. Her own *Alexiad* is a much less fascinating work, but the recovery of the Byzantine power under her father provided her with an important theme.

Procopius' histories

The last, increasingly disastrous, centuries of Byzantine history are recorded by a series of scholarly and interesting historians. Nicetas Choniates, a high imperial official, provides a surprisingly balanced eyewitness account of the siege and capture of Constantinople by the forces of the Fourth Crusade (1202–04). George Acropolites, a leading adviser of the Greek emperors of Nicaea, carries the story from 1203 to the recapture of Constantinople by the Byzantines in 1261. The later 13th and 14th centuries are covered by a succession of writers deeply immersed in contemporary theological disputations. Perhaps the most readable of all Byzantine histories is the largely autobiographical work of the leading politician and emperor John VI (reigned 1347 to 1354), written after his deposition during his years of enforced retirement in a monastery. George Sphrantzes, a close friend of the last emperor, Constantine XI, included in his history an eyewitness account of the siege and capture of Constantinople by Mehmed II in 1453. Two of Sphrantzes' contemporaries chose to write primarily about the Turks. Their methods place them among Renaissance historians. Laonicos Chalcocondyles wrote (in about 1464) an account of the rise of the Turkish state. He did so in the manner of Herodotus, with long digressions on various neighbouring nations. A little later, Critobulos of Imbros, in his account of the Turkish conquest of Constantinople, made Mehmed II his chief hero and modelled his history on Thucydides.

The study of what might be called "historical antiquities" was not much cultivated by Byzantine scholars. The most notable exception was the emperor Constantine VII, but only some fragments of his voluminous collections have survived (dating from about 940 to 959). They include a very interesting account of the various peoples with whom the Byzantines had to deal. Such ancient Greek literature as still survives, including that of all the historians, was preserved by the Byzantine scholars. When, around the year 1400, the teaching of Greek was introduced into Italian universities by Byzantine scholars, they brought also their superior techniques of literary scholarship, transforming thereby the study of Latin authors as well as introducing into western Europe the treasures of Greek literature. One result was the emergence of the new Renaissance historiography.

MUSLIM HISTORIOGRAPHY

Muslim historiography appears to have originally developed independently of European influences. Until the 19th century Muslim writers only very seldom consulted Christian sources and almost never noted events in Christian countries. Fortunately, they displayed at times more curiosity about the non-Muslim peoples of Asia. The first and best history of the Mongol conquests in the first half of the 13th century was the work of a Persian, Joveynī. On a visit to Mongolia in 1252–53, he was able to consult the recently compiled, earliest Mongol narrative (*Secret History of the Mongols*).

The origins of Arabic historiography still remain obscure because of the gap between the legendary traditions of pre-Islāmic Arabia before the start of the Muslim era (AD 622) and the sophisticated and fairly exact chronicles that began to appear in the later 8th and 9th centuries. But while the detailed stages of this development still await reconstruction, the main influences shaping the early Muslim historiography are clear enough. As in the case of the ancient Jews, it was created and perpetuated by religion. Muḥammad (died 632) regarded himself as a successor to a long series of Jewish and Christian prophets, and he made Islām a religion with a strong sense of history. The Qurʾān, Islām's holy book, is full of warnings derived from the lessons of history.

Origins of Arabic historiography

Teachings of Muḥammad not included in the Qurʾān came to be regarded after his death as authoritative tradition left behind by him. All his sayings and actions were therefore carefully treasured and ultimately came to form, in combination with the Qurʾān, the foundation for the body of Muslim law (Sharīʿah), common to all Islāmic communities. These traditions (Ḥadīth) were transmitted orally for several generations, until they were written down in the 8th and 9th centuries. The resultant collections were only partly historical, as myths and inventions crept into them. The scholars who were engaged in preserving and verifying these traditions were chiefly preoccupied with organizing them into legal and theological systems, and they were frequently hostile to the historians. The earliest authoritative life of Muḥammad, written by Ibn Isḥāq (died 768), was attacked by a leading exponent of the legal "traditionist" learning. This confirms the independence of the historical scholars from the theological and legal interests. But both groups shared some common materials, and the strict rules evolved by the legal "traditionists" for recording their sources and tracing a continuous chain of authoritative transmitters of the traditions encouraged similar exact habits in the Muslim historians. The resultant histories were often pedantic, full of unrelated facts, and deficient in reflective comment, though there are some astonishing exceptions, such as the writings of Ibn Khaldūn (1332–1406). But the better Muslim historians scrupulously quoted their authorities and tried to be truthful. This was particularly true of the "classical" school of historians, who were writing at the centre of the ʿAbbāsid caliphate in Iraq in the 9th and 10th centuries. Aṭ-Ṭabarī (died 923), the most authoritative of them all, wrote his "History of Prophets and Kings" as a supplement to his earlier commentary on the Qurʾān, and subsequent Muslim historians were content to follow his reconstruction of the early Islāmic history. The Syrian and Iraqi historiography of the 12th and early 13th centuries is at least as valuable as the Western historical writing of this period, and sometimes it is clearly better.

To orthodox Muslims, the development of the Islāmic community represented a continuous manifestation of God's purpose. Consequently, the recording of the religious progress of the Islāmic society continued to be sacred duty. One of the original features of Muslim historiography is the large amount of attention devoted to the lives of devout men and of scholars. To many Muslim historians, these spiritual and intellectual activities were of much greater importance than the doings of princes and warriors. One of the peculiarities of Muslim historiography was the liking for encyclopaedic dictionaries of famous men. The earliest of these were devoted to the Companions of Muḥammad and to the early transmitters of the Muslim traditions. For a thousand years extremely diverse types of biographical collections have continued to appear in the Muslim world. Those devoted to religious scholars attained a particularly wide diffusion. Saladin (Ṣalāḥ ad-Dīn), who took Jerusalem from the crusaders in 1187 and later opposed the Third Crusade, offered to the Muslim writers the particularly congenial subject of a ruler dominated by a sense of religious duty. A particularly fine example of medieval Muslim historiography is the biography of Saladin by Bahāʾ ad-Dīn (died 1234), which gives

an exceptional insight into Saladin's motives for many of his critical decisions.

The works of Ibn Khaldūn

But the greatest Arab historian and one of the most penetrating thinkers about historiography in any time or place was undoubtedly Ibn Khaldūn. The introduction (*al-Muqaddimah*) to his *Kitāb al-ʿibar*, a universal history (begun in 1375), is, in A.J. Toynbee's judgment (1934), "the greatest work of its kind that has ever yet been created by any mind." Ibn Khaldūn had absorbed all the learning accessible to a Muslim of his time. He was a master of religious learning, an outstanding judge, a writer on logic. He turned a subtle and most disciplined mind to historiography in order to explain his personal tragedy. He had served a succession of rulers in Islāmic Spain and the Maghrib (Northwest Africa) as a general, a politician, and even once as a chief minister, and his activities had always ended in disaster. In order to explain what had gone wrong, he sought to achieve a correct understanding of the forces that governed the societies known to him. He concluded that political stability had become impossible in his native Maghrib, because over centuries economic prosperity had declined excessively and the forces of lawlessness had become too strong.

As a detailed chronicler of events Ibn Khaldūn is not always exact, but, like contemporary historians, he knew how to reconstruct correctly the main trends over several centuries. His ability to formulate general laws that govern the fate of societies and to establish rules for the criticism of sources provided him with an intelligent framework for the correct reconstruction of past history.

Ibn Khaldūn's *Muqaddimah* has survived in at least a score of manuscripts, but he has had no effective influence on Muslim historiography until recently; after his time, as before, the writing of history continued to be a normal feature of Muslim civilization in the more advanced Islāmic societies. In several countries, notably in parts of India, the first works that deserve the name of history appeared only after the Muslim conquest or the conversion to Islām. After the 12th century Arabic ceased to be the main language of Muslim historiography. Distinguished histories were written in Persian in the 13th century, and subsequently Turkish and other vernaculars came to be used by historians in different parts of the Islāmic world. But, in its isolation from non-Muslim influences and its traditional interests, Islāmic historiography underwent no intrinsic change until the 19th century, when it began to be affected by the impact of modern Western civilization.

HISTORIOGRAPHY IN THE EUROPEAN RENAISSANCE

If there is one thing that united the men of the Renaissance, it was the notion of belonging to a new time. Lorenzo Valla, one of the ablest of the early Humanists, in a preliminary draft of his history of King Ferdinand I of Aragon (written in 1445–46), proudly enumerates the modern technical inventions made in recent centuries, and especially near his own day. The sense of the novelty and excellence of their achievements was particularly felt by the men of the Renaissance in connection with their attempts to imitate the works of the ancient Greek and Roman writers and artists. They were not yet claiming that an era of unlimited progress was dawning for mankind—such concepts belong to the 18th century—but the belief in the progressiveness of their own age soon spurred the best Renaissance scholars and artists into achievements that, in some important respects, surpassed their ancient models. This happened in historiography, and especially in the sciences connected with it. The pace of change must not be exaggerated, however. Despite promising beginnings, historiography as a systematic discipline did not emerge during the Renaissance and, in fact, this development did not occur until the 19th century. The reasons for this delay form one of the main problems in any study of historiography between the years 1400 and 1800.

Denigration of the Middle Ages

In the early Renaissance one by-product of the newly won sense of modernity was the tendency to regard the millennium between the collapse of the Roman Empire in the West and the 15th century as an era of prolonged decline. The concept of the Middle Ages was thus introduced for this intervening period. Two very important histories written in the first half of the 15th century deliberately concentrate on the medieval centuries. Their authors were leading Italian Humanists. The first to appear was the *Historiae Florentini populi* ("History of Florence") of Leonardo Bruni, the city's chancellor from 1427 to 1444. The second, the *Historiarum ab inclinatione Romanorum imperii decades* ("Decades"; mainly devoted to Italy), was written by Flavio Biondo, an important papal official. It covered the period from the sack of Rome by Alaric in AD 410 to the writer's own time. The "invention" of the Middle Ages as a separate historical period remains one of the most enduring legacies of Renaissance historiography.

Unlike the medieval historians, the Renaissance Humanists became much more acutely aware of the process of historical change. This was a gradual development. They were trying to understand the ancient writers, whom they were seeking to emulate, and they became increasingly aware of the need to replace these writers in their correct historical setting. When Petrarch (1304–74), the pioneer Italian Humanist, unearthed in 1345 a collection of Cicero's letters, he was shocked to discover that Cicero was not a cloistered scholar of the medieval tradition but a busy politician who wrote his dialogues in moments of banishment from active life. In 1361, in a letter to the Holy Roman emperor Charles IV, Petrarch was able to use his increased familiarity with classical documents to expose a medieval forgery of the Austrian archduke masquerading as a charter of Julius Caesar.

Between about 1440 and his death in 1457, Valla was one of the most influential Humanists. His *Elegantiae linguae latinae* (1444; "Elegancies of the Latin Language") was a treasury of information about correct Latin usages. For Valla the meaning of words was not natural but conventional and historical, because it was derived from changing custom. Thus a sense of ceaseless historical evolution was planted at the very centre of Humanist preoccupations with the recovery, the correction, and the interpretation of ancient texts.

In 1440 Valla's patron, King Alfonso of Naples, at war with the papacy, asked Valla to write a treatise against Pope Eugenius IV. Valla obliged by decisively disproving, on both linguistic and historical grounds, the genuineness of the "Donation of Constantine." From the middle of the 8th century, when this document was probably concocted, it had been used by the popes as one of the weightiest justifications for their claims to secular authority in Italy. Its authenticity had been sometimes questioned in the past by some of the acutest minds, such as Bishop Otto of Freising in the 12th century and Marsilius of Padua in the first half of the 14th century, but it required Valla's expert techniques to dispose of the "Donation" forever. The validity of Valla's methods of historical criticism was at once recognized by at least one other leading Humanist. Biondo wrote the relevant portions of his "Decades" of papal and Italian history between 1440 and 1443, while remaining in the service of the very same Eugenius IV who had been the chief object of Valla's attack. Yet Biondo tacitly accepted Valla's conclusions, and he never mentions the "Donation of Constantine." Biondo's critical outlook found still another expression in his summary dismissal of the fabulous history of Geoffrey of Monmouth. In his copy of Geoffrey he entered only a single note: "I have never come across anything so stuffed with lies and frivolities."

Historical philology

Valla's work on the texts of the New Testament proved in the long run to be one of the most influential applications of the new science of historical philology. His aim was to recover, so far as possible, the original Greek version through the use of the oldest extant manuscripts. He defended these researches by pointing out that he was not correcting the Holy Scriptures but merely the Latin Vulgate translation of St. Jerome that had been adopted by the Catholic Church. The revolutionary nature of Valla's historical approach comes out most strikingly in his comment that "none of the words of Christ have come to us, for Christ spoke in Hebrew and never wrote down anything." The corrections assembled by Valla became generally known when, in 1505, Erasmus published them as *Annotationes* on the New Testament. They provided a model for Erasmus' edition of a Greek New Testament in

1516, from which stem all the new Protestant versions of the 16th century.

The new historical philology was also soon applied to the study of philosophical and legal texts. In this, the most striking progress was made in the second half of the 15th century by Politian, who lectured at Florence, and by his friend Ermolao Barbaro, who taught at Padua. They were inaugurating the history of ideas and of intellectual movements. In his studies of Aristotelian texts, Barbaro insisted on using only the commentators of antiquity. In his lectures and writings (1489–94), Politian tried to reestablish from internal evidence the correct sequence of Aristotelian treatises, and he traced the gradual liberation of Aristotle's thought from the influence of Plato. The meaning of the terms used by Aristotle was rigorously investigated in the light of the linguistic usage of his Greek contemporaries. Politian's ventures into the field of legal texts proved particularly influential. He had at his disposal a very good 6th-century version of the Digest—that is, the section of Justinian's Corpus Juris Civilis (Body of Civil Law) based on the rulings of the Roman jurists. Politian's collation of it with the first printed edition of the Digest (in 1490) formed part of an inquiry into the transmission of the texts of the Roman law during the Middle Ages. Politian's researches stimulated a remarkable school of Humanist jurists, mostly Frenchmen, headed by Guillaume Budé, who published the first historical commentary on the Digest in 1508. In the course of the 16th century, these scholars laid the foundations of a new branch of scholarship, the history of laws and institutions.

The methods of textual criticism used by Politian and his friends were designed to produce definitive editions of classical texts. Politian was aware of the need to establish the correct descent of manuscripts and to disentangle the best textual tradition. In all this he was far ahead of almost all his contemporaries, and he was anticipating the procedures that were systematically adopted for the first time by Karl Lachmann and other German scholars in the 19th century. The historical philology of Politian was a program for the future rather than a dawn of a new era in the editing of classical texts. In contrast to his methods, most of the other Humanist editions of the Latin and Greek classics are very unsatisfactory. This is particularly true of the editions produced between about 1400 and 1550. The reckless emendations of Humanist editors, coupled with the subsequent disappearance of some of the manuscripts used by them, created grave problems for later scholars. Ever since the 17th century the task of the more modern editors has consisted largely in reconstructing, so far as possible, the manuscript versions available before 1400.

Modern historiography was created in the 19th century through a successful combination of the use of narrative sources with every other type of evidence. Some 15th-century Italian Humanists were already aware of these possibilities. The idea of recovering an entire civilization through a systematic collection of all the relics of the past was not alien to them. Biondo used mainly conventional narrative sources for his "Decades" of Italian history, but his description of the city of Rome in antiquity (*Roma instaurata,* 1444–46) was based on a novel combination of the narratives of other historians with a wide range of miscellaneous sources. These included topographical guides, public and private documents, studies of surviving buildings, inscriptions, and coins. But in practice most histories and biographies continued to be written in a conventional way, while the revived study of "antiquities" was cultivated in separation from narrative historiography.

Biondo's description of ancient Rome

Imitation of ancient models is the feature most often stressed in the modern descriptions of Humanist histories. This meant that style mattered at least as much as content and that historical truth might be obscured by literary conventions. On the more positive side, there was the renewed insistence on the choice of definite, clearly delimited subjects and on a more coherent arrangement of material. The abler Humanist historians, however, were also making innovations that bring their practice a little nearer to present notions of writing history.

Several Humanist historians were particularly attracted to the study of the origins of the states about which they were writing. In the 15th century Bruni did this for Florence, and Biondo and Bernardo Giustiniani for Venice, to mention some notable examples. In the 16th and early 17th centuries, French and English scholars inaugurated a critical study of the origins of their national institutions. Humanist historians prided themselves on their critical ability to overthrow the legends in which various countries had concealed their ignorance of their own origins. The incentives to revise the earliest history were often political. Bruni deemed it essential to prove that Florence had not been founded under the tyranny of the Roman emperors but in the time of the free republic. He happened to be right. The Humanist historians were more confident than their ancient predecessors that they could write competent histories of a remote past. In practice they were much less successful in this than they imagined. In dealing with periods before their own time, they usually followed only a restricted number of earlier narratives, though the best of them, such as Bruni and Biondo, displayed in their histories of medieval Italy a novel ingenuity in combining well-chosen sources. Biondo, for example, made effective use of Dante's correspondence.

There was also some modest progress through the better use of documentary sources. This is often far from obvious, because Humanist historians, like their ancient predecessors, do not usually refer to their sources, even when they quote texts verbatim. Hence came Leopold von Ranke's utter misjudgment of the historical value of the *Storia d'Italia* ("History of Italy") of Francesco Guicciardini. Before Ranke's time it was universally accepted as the most authoritative contemporary history of Italy in the years 1494 to 1534. Ranke, who became one of the pioneers of "scientific" history in Germany, first established his reputation in 1824 by his attack on the reliability of Guicciardini. Ranke argued that the statements of that great Florentine statesman were contradicted by documentary evidence and that his history must have been based on unreliable secondary authorities. The discovery in the 20th century of Guicciardini's private archive proved that his history was scrupulously based on original documents of the highest value.

Guicciardini's history of Italy

Guicciardini, in a work that forms the nearest Renaissance parallel to the history of Thucydides, tries to comprehend the succession of tragedies that befell Italy from the start of the French invasions in 1494. This desire to recapture the rational causes of events is one of the most mature features of the best Renaissance historiography.

EARLY MODERN HISTORIOGRAPHY

Italian Humanist historians provided models that could be imitated easily in other countries. Almost everywhere in western and central Europe, local writers were encouraged to produce descriptions and histories of their own lands, intent with patriotic pride. In such countries as Spain and Poland, which had only recently achieved their unity, this was a way of commemorating their newly won cohesion. In the 15th century it was the object of a pioneer work on the earliest antiquities of Spain, the *Paralipomena Hispaniae,* by the Catalan Humanist bishop Joan Margarit i Pau, and of the invaluable *Annales seu cronicae incliti regni Poloniae* ("History of Poland"), by Jan Długosz, which included an exceptionally precise geographic description of his country. In Germany a sense of national identity could be vindicated by Humanist historians striving to minimize the importance of the continued political division of their land. The *Germania* of Tacitus was printed in Germany as early as 1473 and started the fashion of using this collective name for that country. Tacitus called the Germans "the indigenous inhabitants." This was used by a leading patriotic Humanist, Conradus Celtis, as a proof that Germany should be free from all foreign domination. Celtis and his other Humanist contemporaries deliberately hunted for manuscripts of medieval German writers to prove that their country, despite its disunity, could have a national history. Some important masterpieces were recovered, including the histories of Otto of Freising. Celtis' pet project of a description of Germany modelled on Biondo's *Italia illustrata* was carried out in 1530 by Sebastian Münster, and Münster's

fuller *Cosmographia* (1544; "Cosmography"), though purporting to describe the known world, devoted one-half of its 818 pages to "the German nation." There was also a spate of histories of Germany, mostly very laborious and unreflective but incorporating the newly rediscovered medieval narratives and even some documentary sources. Greater originality came only in the wake of the Reformation. The same thing happened in France and in England. In both countries patriotic preoccupations were a leading feature of works written by Humanist historians, and the appearance of Protestantism reinforced in a peculiar way the existing nationalist tendencies.

The influence of the Reformation on historiography must first be discussed at a more universal level. As the philosopher Francis Bacon shrewdly observed, Martin Luther had been obliged "to awake all antiquity and to call former times to his succours . . . so that the ancient authors . . . which had a long time slept in libraries, began generally to be read." This was not because Luther would have regarded himself as a historian. But as early as 1519, in his disputation with Johann Eck, he encountered the assertion that the primacy of the pope was of divine origin. In order to disprove this and to demonstrate that they alone represented the true church, the Protestants had to retell in a new way the entire history of Christianity. In a preface to the *Vitae Romanorum pontificum* ("Lives of the Pontiffs"), published by Robert Barnes in 1535, Luther himself confessed that, although he himself had not originally attacked the papacy with historical arguments,

> now it is a wonderful delight to me to find that others are doing the same thing . . . from history—and it gives me the greatest joy . . . to see . . . that history and Scripture entirely coincide in this respect.

Protestant history

The starting point for the Protestant rewriting of Christian history could best be found in St. Augustine's teachings. The true church, the city of God, had always existed, even though at times it seemed to be overshadowed by the enemies of the divine order. Those enemies were not only the pagans and the heretics, as St. Augustine had believed. In more recent times they had included also the upholders of the papal authority and the persecutors of such medieval true Christians as John Wycliffe (died 1384) and John Hus (died 1415). The writings of Eusebius provided the model for chronicling the sufferings of the faithful until the dawn of freedom for the true church in the 16th century. These views about the correct history of Christianity were presented with exceptional cogency in John Calvin's *Christianiae religionis institutio* (fullest edition 1559; *Institutes of the Christian Religion*) and were shared by most Protestant scholars. The only obvious disagreements arose when Protestants tried to pinpoint the moment at which the church took the fatal turn away from God's true purpose. While the radical sectarians considered that the papacy had always been corrupt, less extremist Protestants were prepared to accept the earlier popes and to argue that the rot set in at some date between the time of Eusebius (died *c.* 340) and the 7th century. The choice of precise date might depend on the national traditions of each country. Thus, Bishop Richard Davies, in his preface to the New Testament in Welsh (1567), treats Pope Gregory the Great (died 604) as a special enemy because Gregory's effort to convert the Anglo-Saxons led ultimately to the subjugation of the autonomous British church.

Historians writing in this spirit were incapable of impartiality. But the historical controversies between the Catholics and the Protestants produced from both sides huge compilations. Their authors were determined to prove their respective cases by a stupendous marshalling of authorities and documentary sources. The habit of giving copious references and long, exact quotations, missing from the Humanist historiography, was reintroduced by the religious controversialists. On the Protestant side, the largest work is the *Ecclesiastica historia,* or the so-called *Centuriae Magdeburgenses* (13 volumes, 1559–74; "Magdeburg Centuries"), retelling the history of the church down to 1200. The Catholic reply, equally huge and graceless, was produced in 12 volumes by Cardinal Baronius. The chief Protestant critic of this work, the great Greek scholar Isaac Casaubon, was astonished by the Cardinal's ignorance of Greek and Hebrew, his gross mistakes, and his boundless credulity.

The narratives of contemporary events written in the 16th and early 17th centuries by the participants in the religious struggles, though equally partisan, include some works of great historical value and high literary merit. The earliest and best German Protestant narrative, that by Johannes Sleidanus, received a grudging tribute from his great opponent, the Holy Roman emperor Charles V, who remarked that "the rogue has certainly known much . . . ; he has either been in our privy council or our Councilors have been traitors." John Foxe's *Book of Martyrs* (1563) contains a great mass of exact information about the persecution of reformed religion in England and Wales during the reign of Mary Tudor, and it has influenced many generations of British Protestants. The achievements of Queen Elizabeth I and the Anglican Church's settlement of her reign found an outstanding defender in William Camden, who was encouraged to write by Elizabeth's leading ministers. In his *Annales Rerum Anglicarum, et Hibernicarum Regnante Elizabetha* ("Annals of Elizabeth's Reign") Camden made excellent use of a mass of official records at his disposal, though his treatment of confidential matters had to be discreet.

Foxe's *Book of Martyrs*

Out of a conflict between Venice and the papacy in the first years of the 17th century was born the *Istoria del concilio tridentino* (1619; *History of the Council of Trent,* 1676) of Fra Paolo Sarpi. A Catholic friar, but a passionate defender of Venetian autonomy, Sarpi drew a dark picture of worldly papal policies and the unscrupulous machinations of the Jesuits. It is a bitter, prejudiced, but splendidly written and well-informed work, which profoundly influenced the anticlerical historians of the 18th century. All these contemporary narratives, however, have one serious limitation. They deal almost exclusively with political events and with changes in ecclesiastical organization. The Protestant schism is treated as merely a revolt against the abuses of the old church, and the deeper reasons for the alienation of the Protestants from the Catholic faith are never explained. Furthermore, these historians, by attributing the origins of the schism almost exclusively to Luther's sudden conflict with the papacy, obscured the existence in the early 16th century of numerous Catholic reformers, whose sole aim was to transform the Catholic Church from within. This one-sided approach to the history of the Reformation was destined to persist for a long time. Two influential histories published in the years 1683–88, one by a great Catholic prelate, Bishop Jacques-Bénigne Bossuet, and the other by Pierre Jurieu, a leading Protestant, still agreed on the same superficial account of the causes of the Reformation.

The rewriting by the Protestants of universal church history naturally involved a drastic revision of the history of the national churches. In Germany, particularly, the history of the church had become inextricably intermixed with the destinies of the German empire. Their hatred of the papacy made the Lutherans visualize the course of German history with unusual clarity. Nobody before them had attempted to impose on that history a single intelligible pattern of any sort. Theirs was bound to be a prejudiced pattern, a story of gradual national disintegration as the result of the successive defeats of the German emperors by the papacy. Johannes Stumpf's tragic chronicle of the Holy Roman emperor Henry IV (published in 1556) treated his struggles with Pope Gregory VII as the beginning of the empire's tribulations. The whole course of German history was retraced in this fashion under the influence of Luther's chief Humanist collaborator, Philipp Melanchthon, in the so-called *Chronicle* of Carion, written in its final versions (1572–73) by Melanchthon's son-in-law, Caspar Peucer.

One of the most novel features of the English Protestant historiography was the reawakening of scholarly interest in the period before the Norman Conquest of England in the 11th century. Matthew Parker, Queen Elizabeth's first archbishop of Canterbury, thought he could discern in the pre-Conquest church elements of true Christianity that were destroyed thereafter and had only been reintroduced by the Protestants. The Anglican Church could be repre-

sented as a return to the traditional practices and beliefs of the early English Christians. Thus the replacement of Latin by English in the Protestant church services could be justified by citing the presence in Anglo-Saxon England of Bibles, liturgies, and devotional literature in the Old English language. Parker and his friend Lord Burghley, Elizabeth's most trusted minister, gathered around them a circle of enthusiastic scholars, whose work preserved most of the important Anglo-Saxon texts as well as of some leading post-Conquest chronicles. Parker's own method of editing texts horrifies modern scholars, but some of the antiquarian works published by members of this group were of high quality. Camden's *Britannia* (first edition 1586, later much enlarged) was a pioneer work on the topography of Roman and early medieval Britain. The edition by Sir Henry Spelman of the records of the pre-Conquest church councils was the first serious attempt to apply to an important type of early sources the best methods of continental scholarship.

Camden's *Britannia*

The growth of a historical outlook can be traced in the 16th century in many diverse fields of learning. For the first time men were realizing that there was a historical side to every branch of knowledge concerned with human affairs. "I have become aware that law books are the products of history," wrote the French legal historian François Baudouin in 1561. In each branch of study there developed a special historical technique particularly appropriate to it. The most sophisticated scholarship was to be found in the field of classical studies. A group of scholars active in the second half of the 16th century were achieving results much superior to the work of the earlier Renaissance classicists. They combined philological expertise with a determination to reach a really adequate understanding of the ancient Greek and Roman civilizations. A few were Italians, such as Carlo Sigonio, but most of the important works were written in France and in the Protestant centres of Switzerland and Holland. As textual critics these scholars were reacting sharply to the earlier, more haphazard, methods of emending and editing classical authors. They were trying to bring the text of one writer after another to a state of near perfection. Some leading ancient historians, such as Tacitus, benefitted greatly from this treatment (edition of Lipsius in 1575). Though their methods do not quite reach the standards of modern scholarship, they anticipate intelligently many of the procedures more systematically adopted in the 19th century. Isaac Casaubon was the first to point out in his edition of Suetonius (1595) that Einhard's 9th-century life of Charlemagne was modelled on the work of that Roman historian. Casaubon's friend Joseph Scaliger renewed the science of classical chronology (1583) and was the first to reconstruct the original Greek *Chronicle* of Eusebius lying behind St. Jerome's Latin translation. Sigonio's pioneer work on the rights and duties of Roman citizens (1560) was later much used by Theodor Mommsen, one of the founders in the 19th century of the modern study of Roman history.

In the course of the 16th century, non-narrative historical work of the highest originality and complexity was being carried on in the legal faculties of French universities. One important stimulus was provided by the existence in France of different legal systems—the uncodified provincial customs in the north and the written law in the south. The latter ultimately derived from the Roman law, and, in the southern French universities, there arose an eager demand for the introduction of the new Italian methods of interpreting the Roman legal texts. Andrea Alciato, a pioneer in the historical treatment of the Roman law, taught at Bourges from 1529 to 1533, and his pupils founded the "Romanist" school of French legal historians.

Legal histories

Important advances were made in the study both of the Roman law and of the origins of the French legal customs, laying virtually the foundations of a new branch of scholarship, the history of law and institutions. François Baudouin published in 1545 the first historical survey of the development of the Roman legal science. The treatise on the custom of Paris by Charles Dumoulin (published 1539–58) resulted from his advocacy of the codification of the northern French legal customs. It was the first scholarly exposition of a body of customary French law derived from feudal practices, and it amounted to a first comprehensive history of European feudalism. It prompted a series of controversial works by a succession of scholars. The Roman, the Germanic, and the Celtic roots of feudalism all found advocates, and the respective claims of Lombard and Frankish texts to provide the best clues were vigorously canvassed. The complexity of the problems presented by the unravelling of the origins of feudalism dawned on scholars for the first time. The most valuable of these attempts to rediscover the "ancient French constitution" were the researches on "the antiquities of France" of Étienne Pasquier (published 1560–1607), which form a basis for all later study of medieval French institutions.

One of the novel features of European civilization in the later 16th and 17th centuries was a secularization of mental interests. Secular learning could now produce ideas more fascinating to intelligent men than theology. History was one of the most popular types of literature sought by a growing reading public. Several treatises on the proper way of writing history appeared in the third quarter of the 16th century. An anthology consisting of 12 such works, including the famous *Methodus* of the French political philosopher Jean Bodin, was published at Basel in 1576. Nearly 100 years later a "Catalogue of the Most Vendible Books in England" (1657) showed that history books constituted a large proportion of the total works published. It has been estimated that between 1460 and 1700 at least 2,500,000 copies of 17 leading ancient historians were published in Europe.

The late 16th century and the 17th witnessed the publication of several great collections of historical materials. The men who undertook these gigantic tasks often were antiquarians accumulating miscellaneous records rather than historians, but they were supplying materials for generations of future historians. Some of the most important publications of sources appeared in France and the Netherlands. Pierre Pithou was a pioneer in editing materials for the history of the Frankish period. The collections of André Duchesne are a vast storehouse of chronicles and other sources for the study of medieval French history. Le Nain de Tillemont edited 20 volumes of records devoted to Roman and church history during the first six centuries of the Christian Era, which a century later furnished one of the principal sources for Edward Gibbon's work *The History of the Decline and Fall of the Roman Empire.* In 1629 a Belgian Jesuit, Jean Bolland, embarked systematically on the editing of records connected with all the saints whose feasts had at any time been celebrated by the church, and this series of publications has been continued to the present day. In the second half of the 17th century, the French Benedictine congregation of Saint-Maur started an immense series of publications commemorating the history of the Benedictines and of other monastic orders. The greatest Maurist scholar, Jean Mabillon, was accepted throughout Europe as the most erudite historian of his time.

In spite of its popularity among an expanding reading public and of the large number of learned editions of materials that it inspired, history was not, for most of the 17th century, one of the sciences that made men proud of living in a modern age. Immense progress was taking place in mathematics, astronomy, and physics. History not only did not seem capable of much further development, but scientifically minded men were beginning to dismiss it as a branch of knowledge that would never be worthy of serious respect. Mabillon's *De Re Diplomatica* (1681) helped to challenge this pessimistic view, but a further century elapsed before history began to be accepted as an authoritative discipline.

Hostility of rulers toward historiography

One major obstacle to the progress of historiography was the hostility of rulers to publications that did not favour their governments. The growth of an influential reading public made rulers increasingly suspicious of historical writings; for example, the censorship exercised by Cosimo I de' Medici, ruler of Florence from 1537 to 1574, precipitated the decline of Florentine historiography. Comparisons with the past also could be invidious. In 1599 Elizabeth I of England censured an author for describing

the deposition of one of her predecessors, Richard II, 200 years earlier. Fear of possible trouble made highly intelligent scholars into one-sided historians. The great jurist Hugo Grotius avoided in his history of the wars of the Dutch against Spain discussions of the religious aspects. Samuel Pufendorf, the historian of the Swedish conquests, carefully left out the internal developments in 17th-century Sweden.

The scholars who in that century were responsible for the great advances in the mathematical sciences were convinced that their achievements would ultimately give mankind a novel mastery over its natural environment. This is particularly true of Francis Bacon and of René Descartes. Their optimism was laying the foundations for a belief in a possibility of continuous progress without which the purposeful and assured historiography of the 19th century would be inconceivable. But the attitude toward history of most of the leading thinkers and scientists of the 17th century was not helpful to its immediate development. Bacon, who wrote a readable and rationally argued biography of King Henry VII of England, attached no importance to accuracy; for example, he antedated Henry's death by a whole year and could not be bothered to undertake any detailed research. Gottfried Wilhelm Leibniz was a great mathematician, but his attempts to apply science to historiography led to mechanistic constructions from which real human beings were largely missing. Numerous influential thinkers were decidedly hostile to history. Descartes, the most eminent of the anti-historical scientists, was not simply disgusted by the unsystematic and imprecise methods of the historians of his time but also doubted whether, strictly speaking, history could be regarded as a branch of knowledge at all. But it is important to remember that much of the 17th-century criticism of history was an attitude of men who simply had other priorities and were concerned to attack doctrines that, for one reason or another, historians seemed to support. In the late 17th century the most successful defenders of history were the members of certain particularly scholarly Catholic orders. Catholicism rested its authority on tradition to a much greater extent than did its Protestant opponents. For Catholic scholars such as Mabillon, the defense of history became really a defense of their religion. They were trying to show that historians were capable of discovering scientifically demonstrable truths. The decisive publication was Mabillon's *De Re Diplomatica* of 1681. A member of a rival order, the Jesuit Daniel van Papebroch, had challenged (in 1675) the authenticity of the oldest charters of two French Benedictine monasteries, Saint-Denis and Corbie. Mabillon applied his powerful critical intelligence not only to vindicating these documents but also to formulating the general rules that must be used to prove the authenticity of medieval records. He illustrated his rules by admirable examples and stated his conclusions with a candor and a common sense that convinced most readers. Mabillon's survey of the tests that must be applied by scholars covered the writing materials, the scripts (thus founding the science of medieval Latin paleography), the seals and other devices of authentication, the official formulas, and the vocabulary used at different periods. Above all, he stressed that the authenticity of a document usually rested not just on isolated details but on consistent correctness of all its features.

The rules of diplomatics

Mabillon was not just a "historical scientist." He had a passionate interest in the past and a vivid historical imagination. He displayed these qualities abundantly in his last and most important work, the *Annales Ordinis s. Benedicti* ("Annals of the Benedictine Order," to 1066). In the *Traité des études monastiques*, (1691; "Manual of Monastic Studies"), he defended the importance of scholarly work as the principal activity of an elite of Benedictine monks. But it would be an anachronism to regard Mabillon and his chief associates as fully comparable to modern historians. They were constrained by the limitations of their time and of their special position as monks. For example, Bernard de Montfaucon, Mabillon's most important successor, is the creator of the science of medieval Greek paleography. But he shares with most of his contemporaries a complete inability to treat the Old Testament as a historical source.

Historical and antiquarian studies developed in 17th-century England in several very distinctive ways. The political struggles and religious controversies of that period made some issues of older English history into matters of immediate practical importance. The other distinctive feature was the delay in the absorption of European continental learning, so that the great progress made in the study of feudal origins in the 16th century began to affect the thinking of English scholars only by about 1625. But there persisted also elements of continuity growing out of earlier Tudor scholarship. The interest in the Anglo-Saxon church and civilization continued to stimulate important editions of records throughout the 17th and early 18th centuries, including, especially, Sir Henry Spelman's edition of the records of church councils and Sir William Dugdale's *Monasticon Anglicanum* (1655–73), which is still valuable today. Another element of continuity with the Tudor period was the perennial interest of the English notables in heraldry, genealogy, and the antiquities of their native regions. Dugdale's *Antiquities of Warwickshire* (1656) set a pattern and a standard for county histories.

Students of English law and institutions, lacking the stimulus that was provided for French lawyers by the diversity of legal systems and by the notable progress in the study of Roman law in that country, continued to ascribe immemorial origins to the common law of England and to approach the development of English institutions in a completely unhistorical spirit. Among the parliamentary opposition to the Stuarts, these attitudes were part of a belief in the "ancient constitution," which these sovereigns were supposed to be defying. Spelman, who was a devout Anglican and a royalist, though a moderate one, was perhaps the first major scholar to break away from this myth. Under the influence of continental publications and correspondents, he accepted that feudal tenure had been introduced into England after the Norman Conquest and that all the English institutions after 1066 must be redefined in feudal terms. But his discoveries were hidden in a dictionary of antiquarian words (*Archaeologus*, vol. 1, 1626; 2 vol. 1664) and made very little impact until some 50 years had elapsed. Spelman had an acute sense of historical development, and he sadly castigated his countrymen for their lack of it in their attitude to parliamentary origins:

Achievement of Sir Henry Spelman

> when States are departed from their original Constitution and that original by tract of time worn out of memory; the succeeding Ages viewing what is past by the present, conceive the former to have been like to that they live in. (*Of Parliaments*, written in about 1640, published 1698.)

His greatest contribution to English history was to grasp that parliaments had developed out of feudal assemblies convoked by the Norman kings and that the Commons were introduced into parliaments subsequently, as a result of the growing prosperity of the lesser landholders. These views first became generally accessible in the 1664 edition of Spelman's dictionary. They were adopted by Robert Brady (in 1681) and by other partisans of the Stuarts and expanded into a Royalist statement of the English past. Violently polemical though this view was, it did at least lay to rest the myth of the immemorial "ancient constitution." The Whig triumph at the Glorious Revolution of 1688, which established a doctrine that the king ruled by parliamentary consent, led to the neglect of these discoveries for much of the 18th century. This was the common fate of much of the research of 17th-century antiquarians, who were very much ahead of their time and were writing for a limited audience. John Aubrey's pioneer description in the 1670s of the prehistoric sites of Avebury and Stonehenge had to wait two centuries for full publication. Even the best of these antiquarians, such as Spelman and Dugdale, were less critical in their handling of the original sources than Mabillon was. Higher standards were reached by a few of their successors in the early 18th century, especially by Thomas Madox, whose *Formulare Anglicanum* (1702) imitated Mabillon by attempting a systematic introduction to English medieval documents. But this did not save Madox from prolonged oblivion. After about 1730 this English tradition of antiquarian scholarship largely ended and remained unfashionable for most of the 18th century.

HISTORIOGRAPHY IN THE AGE OF THE ENLIGHTENMENT

The impulse given to historiography by the Italian Humanists and the religious controversialists had largely spent itself by about 1715. Men knew again how to write rationally satisfying contemporary histories, though often it needed courage to do so. Much less progress had been achieved in reconstructing the more distant past. Impressive collections of historical materials were being accumulated, but most scholars still lacked the capacity to rethink the thoughts of past generations and thus really to understand them. Mabillon could write with insight about early Benedictine history, as he possessed both sympathy with the subject and adequate technical expertise, but he was exceptional. Spelman had grasped that a particular society would be molded in a peculiar way by its institutions. He could not reconstruct and explain the gradual changes from one set of institutions to a later one, but he was aware of the problem.

Judged by the quality of its historical output, the 18th century was not, on the whole, an age of successful historians, but some of the defects of earlier historiography were beginning to be overcome. There were also losses, however, for some of the achievements of the preceding period were in danger of being forgotten. In the leading countries of western Europe, religious controversies were becoming less important, and a massive secularization of interests took place, which affected even ecclesiastical scholars. The French Maurists continued until 1790 to publish imposing historical collections, but their choice of subjects was determined much less than in the time of Mabillon by religious priorities. The greatest Italian ecclesiastical disciple of Mabillon was Ludovico Antonio Muratori, a social reformer. In a divided country like Italy, the best way of expressing his patriotism lay in reminding Italians of the former greatness of their country. Muratori spent much of his long life on his editions of Italian medieval sources.

The nationalist motivation shown by Muratori was peculiar to Italy and also to parts of Germany, another divided country. Elsewhere in Europe there was a danger that, as men lost interest in constitutional or religious disputes that might be settled by appeals to the past, they might turn away altogether from history or at least neglect long stretches of it. This did happen to some extent in the 18th century. Some of the radical French reformers, such as Jean Le Rond d'Alembert, one of the main inspirers in the 1750s of the French *Encyclopédie,* wanted to jettison completely much of the past. The Marquis de Condorcet, an early prophet of the doctrine of endless progress of mankind and a pioneer historian of European civilization, was a prominent member of a French parliamentary commission that in 1792–93 deliberately destroyed some of the royal records as comprising relics of past servitude.

During much of the 18th century it was safer and easier to publish controversial works of history than it had been in the past. The point is important, as without this greater freedom, the peculiarly radical "philosophical" historiography, so typical of that century, would have been inconceivable. In Italy such writing was still dangerous. Pietro Giannone, the author of an anticlerical history of Naples (1723), was tracked down by the Inquisition and spent 12 years in prison, where he died in 1748. Even the great Muratori, who tried to help Giannone, came into danger of having some of his works banned and had to be rescued by the personal intervention of Pope Benedict XIV. In France, Louis XIV in 1714 imprisoned Nicolas Fréret in the Bastille for alleging (correctly) that the Franks were originally a confederacy of German tribes and not descendants of more illustrious ancestors. Under the successors of Louis, nothing quite so absurd happened again, but critics of the government or the church were often in trouble. Great Britain, Holland, Switzerland, and parts of Germany, on the other hand, provided safe oases where most things could be published. It was no accident that the most independent and historically minded group of German professors should have congregated at the University of Göttingen, founded in 1734, in the Hanoverian territory of the kings of Great Britain.

"Philosophical" historiography

A real renewal of historiography in the 18th century could only come if fresh reasons were discovered for making it again worthwhile. Nationalism could supply one such motive; but this only became decisively influential in the 19th century. An alternative was a historiography inspired by the progress in the natural sciences and based on formulating the general rules governing the development of human societies. The chief features of this "new" historiography were a sense of the unity of all human history, including an interest in the continents outside Europe; a capacity for bold generalizations about the salient features of particular periods or societies; and a preference for topics connected with the progress of human civilization. Condorcet's historical sketch of the progress of the human mind, written in 1794, subdivided all known history into nine periods, each starting with some great invention or with geographical discoveries.

The shortcomings of this "rationalistic" historiography have been rehearsed often enough. For many of its writers it was primarily a weapon of propaganda against their enemies in church and state. Their redeeming virtue was the fearlessly critical attitude to all existing authorities, however august or sacred. The vast scale of their generalizations often precluded any detailed research. This was particularly true of the attempts to write histories of civilization, as the existing collections of printed materials did not cater for such interests, while systematic research in archives was seldom possible in the 18th century. In preparing his pioneer essay on the history of civilization, covering the millennium from the Carolingians to Louis XIV (*Essai sur les moeurs et l'esprit des nations,* 1745–53), the French author Voltaire had to collect bits and pieces from most diverse sources.

One of the most valuable achievements of the thinkers of the 18th century was their capacity to study particular societies as coherent units and to formulate the theory that the various aspects of each society's life were closely interrelated. This was not an entirely novel idea, but it first became commonly accepted during this period. Nor were all its adherents anticlericals. Giambattista Vico, a Neapolitan Catholic, was ahead of his contemporaries in his particularly subtle sense of the complex influences by which one phase of society gives place to another. In his reconstruction of these transitions during the early stages of Roman history, he makes no clear lines between periods. His countryman Giannone explains in his autobiography that he had studied Roman law not for its own sake but in order to understand the changes in the society of the Roman Empire. The French philosopher Montesquieu, who owed much to Giannone, was not really a historian, but he displays an acute sense of historical realities. His *De l'esprit des lois* (1748; *The Spirit of Laws*), more than any other book, accustomed his contemporaries to ponder the complex factors that shaped each society. It inspired Gibbon's definition of the kind of history he wanted to write. It was to be a "history related to and explained by the social institutions in which it is contained."

This ideal was realized in Gibbon's *History of the Decline and Fall of the Roman Empire* (1776–88), one of the masterpieces of "philosophical" historiography. Gibbon was preoccupied above all with the problem of human progress. The belief that continuous progress was possible for mankind had been publicly formulated in the mid-18th century by Anne-Robert-Jacques Turgot in France and by Adam Smith in Scotland, independently, it seems, of each other. Gibbon had read works and known scholars influenced by both these thinkers. A belief in continuous progress would confer a new purposefulness on the study of the entire course of human history and could justify a lengthy account of what otherwise might have seemed very obscure stretches of the past. Such a justification was to inspire most of the historiography of the 19th century. But the problem of progress had a special urgency for Gibbon's generation, which worried at the thought that their own enlightened civilization might also subsequently collapse. By unravelling the causes of the decline of the Roman Empire, Gibbon was determined to show that the Europe of his own day had attained a much superior degree of development and was immune from the fate of the ancient world.

Gibbon's *Decline and Fall*

In the 18th century, historiography was still only very

rarely connected with the universities; and thus, except in such isolated places as Göttingen in Germany, no continuous schools of history could develop. Some of the most important achievements of the 18th-century historians meant much less to their contemporaries than to their successors in the 19th century. Gibbon was a pioneer in utilizing in a "rationalist" history the vast materials accumulated by generations of erudite antiquarians, but he had no immediate followers. The German archaeologist Johann Joachim Winckelmann tried to revive the true understanding of Greek sculpture and to make the history of art into something more than just the biographies of artists, but his work bore little fruit until the next century. The saddest fate was that of Vico's work. He was hardly ever read before the 19th century, when he at last influenced Barthold Georg Niebuhr and the rest of the German historical school, while Jules Michelet's rediscovery of Vico in 1824 started a new era in French writing on the Middle Ages.

HISTORIOGRAPHY IN THE 19TH AND 20TH CENTURIES

From the early 19th century, historiography began to develop in a radically different way. The decisive changes occurred among the German historians, largely through a reaction to the French Revolution and to a temporary subjugation of their country by Napoleon. Organized teaching of history in schools and universities became a matter of national importance, first in Prussia and then in other parts of Germany. As universal education spread to most European countries in the course of the 19th century, history was accepted everywhere as a necessary subject in schools. For the first time the bulk of historical writing came to be done by professional historians, for whom it became a condition of securing academic appointments or of consolidating their standings as university teachers. Historiography eventually became a continuously cooperative venture, where the achievements of past historians could be used systematically by their successors. But the growth of specialization and the bewildering number of types of works that came to be published constituted a new danger. In the past, important discoveries were frequently lost through lack of interest. But, by the second half of the 20th century, discoveries were in danger of being simply overlooked amid the flood of publications.

Growth of specialization

Another great change lay in the growth of intellectual freedom. Free expression of independent or unorthodox ideas had become dangerous during the French Revolution and under Napoleon, both in the territories controlled by the French and, by way of frightened reaction, in the lands of their unconquered opponents. After 1815 conditions for freer historiography improved gradually in much of Europe. Charles Darwin's *Origin of Species* (1859), which put forth a theory of evolution at first unacceptable to church authorities, probably could not have been published with the same impunity any earlier.

One feature of the growing tolerance of governments toward historiography was the gradual creation of public archives, such as the British Public Record Office in London, created in 1838, and the freer opening of the collections already in existence. Even the papacy accepted these changes, and Pope Leo XIII opened up the papal archive in 1883 as part of a deliberate new policy of encouraging historical study of Catholicism. For the first time historiography came to be based largely on unpublished records, and scholars were tempted into excessive reliance on original documents while unduly neglecting the older types of narrative sources.

In the 20th century some grievous threats to the persistence of free scholarship recurred, and historiography suffered with other branches of humane studies. The establishment of a Communist regime in Russia led, at first, to the rejection of most pre-1917 history as a fit subject for schools and universities. This decision was reversed in the 1930s, and from 1945 Communist countries were encouraging a form of historiography especially concerned with economic history and the class struggles of the past. There was also an enthusiastic interest in the material remains of past ages, leading to an impressive development of archaeology, particularly in Poland. The rise of dictatorships in Italy and Germany had disastrous effects on historiography in those countries, and recovery after World War II was only gradual.

Judged merely by the number of "practicing" historians and of their publications, historiography seemed in a very flourishing state in the 1970s. Its European traditions had spread to all the other continents and were largely accepted in all non-Communist countries.

The *Introduction aux études historiques* (*Introduction to the Study of History*) of Charles V. Langlois and Charles Seignobos (1898), supplemented by critical comments of another outstanding French historian, Ferdinand Lot (in *Le Moyen Age,* 1898), provides an excellent starting point for the discussion of modern historical methods. History is an autonomous branch of learning, and some of its methods may be unique. Historians should not try to formulate general laws; their branch of learning merely "aims at explaining reality." Langlois and Seignobos particularly stress that history is not a science of observation but a science of reasoning how to extract from imperfect documentary or narrative records some glimpses of what actually happened.

The historian's task

A historian has to subject his sources to a whole series of preliminary investigations. First comes "external criticism," aimed at determining whether the sources are appropriate and adequate for the particular task in hand. The provenance, date, and authenticity of each source must be established by using the techniques of diplomatic, the detailed study and assessment of documents, and of paleography, the study of ancient handwriting, and of other auxiliary sciences that were elaborated after the 17th century. In France a special institution for teaching some of these techniques, the École des Chartes, was created in 1821. The first specialized seminar for instruction in these subjects was established in 1854 at Vienna by Theodor von Sickel, one of the greatest medievalists of the 19th century, and it was gradually imitated by leading German universities. One of the most important critical refinements introduced in the course of the 19th century was the improved handling of narrative sources brought about by seeking to discover the literary sources that lay behind them. Leopold von Ranke, one of the foremost German historians, who began his career as a teacher of classics, was gradually attracted to history through a desire to understand better the sources of the Greek and Latin authors whom he was expounding. In the later decades of the 19th century, such a quest became a normal feature of historical scholarship.

Once a historian has decided, through the application of "external criticism," on the sources that are relevant to his purpose, he must next, by "internal criticism," make sure that he fully understands what he has selected. German classical philologists were the first to bring these latter investigations to a high degree of perfection. Karl Lachmann, an editor of the Latin poets, is justly regarded as the creator of modern textual criticism in its most rigorous forms, and historians gradually adopted similar methods. The language of the sources must be understood, corruptions in the text must be eliminated, and the historian must, as accurately as possible, penetrate the minds of the authors with whom he is dealing.

All these critical operations on the sources are merely preliminaries, and the work of the historian proper only starts when he attempts a synthesis of his materials. F. Lot stresses that in this qualities other than the erudite skills come into play. There must be sympathy with the subjects under study, for without it there can be no imaginative insight into the past. Ideally, a historian must display capacities akin to those of a poet or an artist.

Such a quality was, by and large, lacking in the work of the historians of the Enlightenment, who had been unable to achieve imaginative insight into civilizations very different from their own. The greatest shortcoming of Gibbon was his temperamental inability to appreciate religion. The new historiography of the 19th century was created chiefly by Germans, who, through a reaction to the ungodly and cosmopolitan Enlightenment, were endowed to excess with a passion for extolling the unique nature of their fatherland and for tracing the roots of this

uniqueness through the whole course of German history. These developments in German historiography can be traced back to some strands of German thought in the 18th century, especially to some features of the writings of Johann Gottfried von Herder. He denied that the purpose of history was to provide a bird's-eye view of the progress of the human mind. It was, rather, to reconstruct history as it had been, which means that all countries and periods are equally deserving of study. This view anticipated Ranke's oft-quoted aim to describe what has actually happened and his conviction that the description of all human history displays the workings of God's providence. The disasters inflicted upon Germany by Napoleon brought forth a patriotic school of historians whose urgent task it became to propagate these views as a means of restoring German independence. The centre of this movement was in Prussia, at the newly founded University of Berlin (1809). Wilhelm von Humboldt, its effective founder, believed that the task of the historian lay in discovering the ideas behind the facts. The concepts that had special validity for him were ideas of religion and of a national state. The German historical school prided itself on the scientific precision of its methods, on its determination to get all the details right, and on the scrupulous quotation of sources. This display of exact scholarship represented a great gain for historical sciences, but its chief purpose was to convince the reader. Yet these German historians were fundamentally inspired by a prejudiced, arbitrary set of assumptions. It is particularly difficult to detect Ranke's hidden bias, as he made a parade of refusing to pass judgments on the past. His preference for the study of foreign relations between states and his treatment of states as natural entities with a right to fulfill their individual destinies justified the successes of Prussia. The defeat in 1848 of the German aspirations to national unity inspired his pupil Wilhelm von Giesebrecht to write the history of the medieval German empire to remind his countrymen of their past glories. When German unification was achieved in 1871, Giesebrecht doubted whether there was any need to bring out any further volumes of his great work. But many German historians, having contributed mightily to the unification of Germany, continued to describe complacently the triumphs of the Bismarckian state. This was one of the purposes of the school of historical economists led by Gustav von Schmoller. There were some dissenting voices. Theodor Mommsen, the greatest historian of antiquity produced by the 19th century, deplored the tendency of his countrymen to worship state power. Friedrich Meinecke, a leading German historian of political ideas, who until 1914 accepted the ordinary nationalistic assumptions of his countrymen, gradually entirely changed his views and, after the defeat of Germany in two world wars, pleaded in his *Deutsche Katastrophe* (1946; *The German Catastrophe*) for a historiography concerned with the higher values of general civilization. Among the German historians, particularly striking progress was achieved in medieval studies. Meanwhile, attempts at imaginative reconstructions of the past were being made in other countries of western Europe. Jules Michelet wrote in 1833–43 the first history of medieval France based on the French national archives, of which he was at that time keeper. Macaulay's *History of England* (1848–61), covering chiefly the years 1685–1702, represented again a remarkable though prejudiced attempt to relive the past.

German patriotic historians

German scholarly techniques and the methods of German historical teaching spread to other countries in the course of the later 19th century, though it is important to note that until 1914 a significant proportion of leading historians from states outside Germany spent some time in that country. This is particularly true of some of the greatest Russian scholars, such as M.I. Rostovtzeff, one of the most important modern historians of antiquity. In England, William Stubbs, though self-taught, applied the results of German scholarship to the reconstruction of English medieval history. Gabriel Monod, who had studied in Germany, was prominent in introducing more scientific techniques into medieval French historiography, and he founded in 1876 the *Revue Historique* as the main organ of French historical scholarship. A succession of American students went to Germany, and some, on their return home, reorganized historical studies. Measured by the sheer bulk of publications, the amount of American history written since the 18th century is probably greater than that of any other modern nation. But apart from editions of sources, very few works on American history published before about 1900 are of much practical use today. The most influential pioneer in organizing scientific historiography was Herbert Baxter Adams, who between 1876 and his death in 1901 made the Johns Hopkins University at Baltimore into the foremost American centre of historical studies. He was also one of the founders of the American Historical Association in 1884 and played a large part in successfully launching the *American Historical Review* in 1895 as the main organ of historical scholarship. Some of Adams' pupils became great scholars in various fields of general history. Charles Homer Haskins' works on Norman institutions and on science and culture in the 12th and 13th centuries made him one of the foremost medievalists of the 20th century. But a movement for creating a purely American history was launched in 1893 by another of Adams' pupils, Frederick Jackson Turner, who inaugurated a "progressive" school of historians through his conviction that the fundamental fact of American history down to 1890 was the settlement of a continent. In Turner's eyes the main theme of American history in the 19th century was the conflict between the patrician and capitalist groups of the Eastern Seaboard and the needs of the new settlers in the Middle West. Charles A. Beard inaugurated by his *Economic Interpretation of the American Constitution* (1913) an attempt to rewrite the entire history of the U.S. in terms of conflicts between different groups of economic interests. The weakness of this type of historiography was that it encouraged an excessive parochialism. After 1945 the "progressive" historians came under fire both from more conservative scholars who preferred to stress elements of common tradition and purpose in American development and from the historians of the "new left." In the 1960s and 1970s the close connection between writings on American history and the active political life was infusing great variety and vitality into its historiography, though making it perhaps too susceptible to rapidly changing external pressures.

The work of Henry Baxter Adams

(E.B.Fr./Ed.)

Methodology of historiography

The methodology of history does not differ in broadest outline from that of other disciplines in its regard for existing knowledge, its search for new and relevant data, and its creation of hypotheses. It is the same for all historical writing, success depending on skill and experience; and division of the past on temporal or topical lines merely reflects the human limitations of historians. Although historical methodology has four facets, the more skilled the historian the less he gives them conscious consideration; and any historian is likely to be concerned with two or more concurrently. The four facets are heuristic, knowledge of current interpretation, research, and writing.

The first two may be briefly considered. Heuristic has been adopted as a convenient term for the technique of investigation that can be acquired solely by practice and experience. In the case of the historian it embraces such things as knowledge of manuscript collections, methods of card indexing and classifying material, and knowledge of bibliography. It underlies other aspects of methodology as in knowledge of the capabilities of historians working in the same and similar fields or in the power of dealing expeditiously with documentary material. The necessity for knowledge of current interpretation is based on the working principle that inquiry proceeds from the known to the unknown; and the historian has to be well acquainted with existing work in his own field, in contiguous historical fields and in allied disciplines. The work in each case consists of both "fact" and interpretation, and the amount the historian accepts will vary. In his own field he will normally not accept facts, and certainly not current interpretation, on trust; in contiguous historical fields he will accept facts and current interpretation by experts in those

fields, but qualified by heuristic and his general historical knowledge; in allied disciplines, such as anthropology, economics, geography, natural science, philology, psychology, sociology, he must unless there is strong evidence to the contrary presume the technical skill and intellectual honesty of scholars in those fields. There is, of course, no reason why a historian cannot be reasonably versed in one or more of these and other disciplines, and should the nature of his enquiry demand it he must be.

Historical research is the term applied to the work necessary for the establishing of occurrences, happenings, or events in the field with which the historian is concerned. Knowledge of these is entirely dependent on the transmission of information from those living at the time, and this information forms what is known as the source material for the particular period or topic. The occurrences themselves can never be experienced by the historian, and what he has at his disposal are either accounts of occurrences as seen by contemporaries or something, be it verbal, written, or material, that is the end product of an occurrence. These accounts or end products have been variously termed relics, tracks, or traces of the occurrences that gave rise to them; and from them the historian can, with varying degrees of certainty, deduce the occurrences. The traces are thus the "facts" of history, the actual occurrences deductions from the facts; and historical research is concerned with the discovery of relevant traces and with deduction from those traces insofar as this will aid the search for further relevant traces.

SOURCE MATERIAL

Source material falls into three groups which can be differentiated as written, material, and traditional. Written source material has two subdivisions, literary, sometimes called subjective, and official. The first consists of events as seen through the eyes of an individual and therefore as interpreted by him, normally entailing selection of occurrences or attribution of motive. The second subdivision, the official, consists of records produced in transacting business at any level from individual to international. The information given is basically in statement form, impersonal, and containing only the most superficial suggestions of causation and motivation. In practice the boundary between literary and official sources is blurred and a document may contain elements of both. The second main division, material source material, consists of objects that have resulted from activities of human beings in the past. The third group, traditional source material, covers what is handed on verbally or as practices, although later generations may commit such things to writing. Obvious examples are archaic forms, traditional practices, nursery rhymes, folklore, and place names. Comparison with parallel source material and knowledge of current interpretation will normally show the historian whether his particular source can be presumed true, partially true, or faked. If true or partially true allowance has to be made for the subjective element in literary and some traditional sources and for the difficulty of reconstructing the events themselves from the traces surviving in official, material, or other traditional sources.

The classification of source material is essentially pragmatic, based on the differing techniques required in handling sources of the different groups: an inscribed tombstone, for example, can be either a written or a material source depending on whether the historian's concern is with the content of the inscription or with the stone. Specialized training in what are sometimes known as ancillary disciplines may, depending on the nature of his investigation, be necessary for the historian. The most important of these are archaeology, bibliography, chronology, diplomatics, epigraphy, genealogy, paleography, sigillography, and textual criticism. It need hardly be said that the historian must have competence in the languages used in his source material. Many historians give part of their time to the editing of source material. This is not historical writing but is of use to other historians in the same field. The collection of facts as an end in itself is, however, antiquarianism not history, and the essential end product of historical investigation is the historian's own writing.

USING SOURCE MATERIAL

The question of what history is belongs to the philosophy rather than the methodology of history. The word history itself is used ambiguously to describe both the past and what is written about the past; but it is this second meaning that is relevant to the working definition that history is the past experience of society. For what reasons society may wish to utilize its past experience is not the concern of the historian, whose task is to make available to society that past experience and to record it for future reference. An individual utilizing his own past experience has to recall the significant elements of that experience with accuracy and establish their causal and chronological relationships. The historian behaves similarly concerning the past experiences of society; but the reconstruction of events from traces, the selection of those relevant to his task, and the establishing of relationships allow a varying freedom of choice by the historian, which thus introduces the subjective element of the historian's personality. This cannot be eliminated from historical writing, and the historian's aim is to make the margin of intellectual error as small as possible. The handling of source material demands only care and technical competence, and it is mainly in the construction of hypotheses and in the establishing of relationships that this intellectual error can enter. A check is provided by the opinions of other historians working in the same field. His work will, if accepted, become part of current interpretation, sometimes described as accepted history but, as with all current interpretation, subject to revision by himself or others.

Historical methodology became more clearly formulated during the 19th and 20th centuries, but there have been historians at times long past whose work can be judged by present-day standards. There are, however, certain important differences between present methodology and the general run of past methodology. Much medieval writing, for example, bows to precedent in literary sources and in current interpretation, and uncritical acceptance of an earlier writer's work can occur century after century. The comparative neglect of official sources by the majority of European historians before the 19th century gave no corrective to literary sources. The greatest impediment to the development of modern methodology lay, however, in the varying concepts of history, some of which survive today. The concept of history as a form of literature made it a type of imaginative art on which judgment was passed on grounds of elegance rather than accuracy. Closely allied with this is the "ethical" concept of history whereby historical writing became a series of value judgments on individuals and actions. The converse of this was the impossible "objective" or "scientific" history of the later 19th century, though it did popularize the concept of research and developed the ancillary disciplines. The use of history for propaganda purposes is in its crudest form virtually a branch of fiction and thus independent of research; in its more subtle forms it can encourage accuracy in research, but it will encourage also the suppression of inconvenient traces and intellectual dishonesty in the elucidation of relationships. In this it indicates one of the main impediments to methodological and historical development: the holding by the historian of *a priori* theories or laws to which all events and relationships must conform, whether it be the theory of divine intervention in human affairs favoured in medieval times or the Marxist theory current over much of the modern world.

(Ed.)

Ancillary fields

ARCHAEOLOGY

The word archaeology comes from the Greek *archaia* ("ancient things"), and *logos* ("theory" or "science"). It has been used in varying ways, but from the late 18th century onward has come to mean that branch of learning that studies the material remains of man's past. This includes man's artifacts, from the very earliest stone tools, of perhaps 2,000,000 years ago, to the man-made objects that are buried or thrown away at the present day: everything made by human beings—from simple tools to complex

machines, from the earliest houses and temples and tombs to palaces, cathedrals, and pyramids.

The archaeologist is first a descriptive worker: he has to describe, classify, and analyze the artifacts he studies. An adequate and objective taxonomy is the basis of all archaeology, and many good archaeologists spend their lives in this activity of description and classification. But the main aim of the archaeologist is to place the material remains in historical contexts, to supplement what may be known from written sources, and, thus, to increase understanding of the past. Ultimately, then, the archaeologist is a historian: his aim is the interpretive description of the past of man.

Increasingly, many scientific techniques are used by the archaeologist, and he uses the scientific expertise of many persons who are not archaeologists in his work. The artifacts he studies must often be studied in their environmental contexts; and botanists, zoologists, soil scientists, and geologists may be brought in to identify and describe plants, animals, soils, and rocks. Radioactive carbon dating, which has revolutionized much of archaeological chronology, is a by-product of research in atomic physics. But although archaeology uses extensively the methods, techniques, and results of the physical and biological sciences, it is not a natural science; some consider it a discipline that is half science and half humanity. Perhaps it is more accurate to say that the archaeologist is first a craftsman, practicing many specialized crafts (of which excavation is the most familiar to the general public), and then a historian.

The nature of archaeological work

The justification for this work is the justification of all historical scholarship, namely, to enrich the present by our knowledge of the experiences and achievements of our predecessors. Because it concerns itself with things people have made, the most direct findings of archaeology bear on the history of art and technology; but, by inference it also yields information about the society, religion, and economy of the people who created the artifacts. Also, it may bring to light and interpret previously unknown written documents, providing an even more certain form of evidence about the past.

But no one archaeologist can cover the whole range of man's history, and there are many branches of archaeology divided by geographical areas (such as classical archaeology, the archaeology of ancient Greece and Rome; or Egyptology, the archaeology of ancient Egypt) or by periods (such as medieval archaeology and industrial archaeology). Writing began 5,000 years ago in Mesopotamia and Egypt; its beginnings were somewhat later in India and China, and later still in Europe. The aspect of archaeology that deals with the past of man before he learned to write has, since the middle of the 19th century, been referred to as prehistoric archaeology, or prehistory. In prehistory the archaeologist is paramount, for here the only sources are material and environmental.

The scope of this article is to describe briefly how archaeology came into existence as a learned discipline; how the archaeologist works in the field, museum, laboratory, and study; and how he assesses and interprets his evidence and transmutes it into history.

History of archaeology. No doubt there have always been people who were interested in the material remains of the past, but archaeology as a discipline has its earliest origins in 15th- and 16th-century Europe, when the Renaissance Humanists looked back upon the glories of Greece and Rome. Popes, cardinals, and noblemen in Italy in the 16th century began to collect antiquities and to sponsor excavations to find more works of ancient art. These collectors were imitated by others in northern Europe who were similarly interested in antique culture. All this activity, however, was still not archaeology in the strict sense. It was more like what would be called art collecting today.

The Mediterranean and the Middle East. Archaeology proper began with an interest in the Greeks and Romans and first developed in 18th-century Italy with the excavations of the Roman cities of Pompeii and Herculaneum. Classical archaeology was established on a more scientific basis by the work of Heinrich Schliemann, who investigated the origins of Greek civilization at Troy and Mycenae in the 1870s; of M.A. Biliotti at Rhodes in this same period; of the German Archaeological Institute under Ernst Curtius at Olympia from 1875 to 1881; and of Alexander Conze at Samothrace in 1873 and 1875. Conze was the first person to include photographs in the publication of his report. Schliemann had intended to dig in Crete but did not do so, and it was left to Arthur Evans to begin work at Knossos in 1900 and to discover the Minoan civilization, ancestor of classical Greece.

Egyptian archaeology began with Napoleon's invasion of Egypt in 1798. He brought with him scholars who set to work recording the archaeological remains of the country. The results of their work were published in the *Description de l'Égypte* (1808–25). As a result of discoveries made by this expedition, Jean-François Champollion was able to decipher ancient Egyptian writing for the first time in 1822. This decipherment, which enabled scholars to read the numerous writings left by the Egyptians, was the first great step forward in Egyptian archaeology. The demand for Egyptian antiquities led to organized tomb robbing by men such as Giovanni Battista Belzoni. A new era in systematic and controlled archaeological research began with the Frenchman Auguste Mariette, who also founded the Egyptian Museum at Cairo. The British archaeologist Flinders Petrie, who began work in Egypt in 1880, made great discoveries there and in Palestine during his long lifetime. Petrie developed a systematic method of excavation, the principles of which he summarized in *Methods and Aims in Archaeology* (1904). It was left to Howard Carter and Lord Carnarvon to make the most spectacular discovery in Egyptian archaeology, that of the tomb of Tutankhamen in 1922.

The development of Egyptology

Mesopotamian archaeology also began with hectic digging into mounds in the hopes of finding treasure and works of art, but gradually these gave way in the 1840s to planned digs such as those of the Frenchman Paul-Émile Botta at Nineveh and Khorsabad, and the Englishman Austen Henry Layard at Nimrud, Kuyunjik, Nabī Yūnus, and other sites. Layard's popular account of his excavations, *Nineveh and Its Remains* (1849), became the earliest and one of the most successful archaeological best sellers. In 1846 Henry Creswicke Rawlinson became the first man to decipher the Mesopotamian cuneiform writing. Toward the end of the 19th century, systematic excavation revealed a previously unknown people, the Sumerians, who had lived in Mesopotamia before the Babylonians and Assyrians. The most impressive Sumerian excavation was that of the Royal Tombs at Ur by Leonard Woolley in 1926.

First steps to archaeology. The development of scientific archaeology in 19th-century Europe from the antiquarianism and treasure collecting of the previous three centuries was due to three things: a geological revolution, an antiquarian revolution, and the propagation of the doctrine of evolution. Geology was revolutionized in the early 19th century with the discovery and demonstration of the principles of uniformitarian stratigraphy (which determines the age of fossil remains by the stratum they occupy below the earth) by men like William Smith, Georges Cuvier, and Charles Lyell. Lyell, in his *Principles of Geology* (1830–33), popularized this new system and paved the way for the acceptance of the great antiquity of man. Charles Darwin regarded Lyell's *Principles* as one of the two germinal works in the formation of his own ideas on evolution. Early stone tools had first been discovered and identified in the late 18th century. The location of these genuine artifacts in certain strata demonstrated, on the principle of uniformitarianism, that man had been in existence long before 4004 BC, the date calculated according to the chronology of the Bible. The discoveries of Jacques Boucher de Perthes in the Somme Valley in France, and of William Pengelly in the caves of South Devon in England, were used to demonstrate the antiquity of man in 1859, the same year that saw the publication of Darwin's revolutionary *Origin of Species.* Approximate dates for the Paleolithic Period (Old Stone Age) of the prehistoric past were thus established, although the expression "Palaeolithic" was not used until John Lubbock coined it in his book *Pre-historic Times* (1865).

The age of the human species

Half a century before this, Scandinavian archaeologists had created a revolution in antiquarian thought by postulating, on archaeological grounds, successive technological stages in man's past. C.J. Thomsen classified the material in the Copenhagen Museum, opened to the public in 1819, on the basis of three successive ages of Stone, Bronze, and Iron. His pupil and successor, J.J.A. Worsaae, showed the correctness of this museum arrangement by observed stratigraphy in the Danish peat bogs and barrows (funerary mounds). Low lake levels in Switzerland in the mid-1850s permitted the excavation of the prehistoric Swiss lake dwellings, and here again the theory of a succession of technological stages was confirmed.

Darwin's *Origin of Species* implied a long past for man, and the acceptance of the idea of human evolution in the last four decades of the 19th century created a climate of thought in which archaeology flourished and that led to great advances in the unfolding of the full story of man's development.

In his *Pre-historic Times,* Lubbock expanded the three-age system of Thomsen and Worsaae to a four-age system, dividing the Stone Age into Old and New periods (Paleolithic and Neolithic). In the last quarter of the 19th century remarkable Paleolithic discoveries were made in France and Spain; these included the discovery and authentication of actual works of sculpture and cave paintings from the Upper (later) Paleolithic Period (*c.* 30,000–*c.* 10,000 BC). When Marcellino de Sautuola discovered the cave paintings at Altamira, Spain (1875–80), most experts refused to believe they were Paleolithic; but after similar discoveries at Les Eyzies in France around 1900, they were accepted as such and were recognized as one of the most surprising and exciting archaeological discoveries. A succession of similar finds has continued in the 20th century. The most famous of these was at Lascaux, France, in 1940.

During the last quarter of the 19th century, Gen. A.H. Pitt-Rivers' excavations of prehistoric and Roman sites at Cranborne Chase, Dorset, laid the foundations of modern scientific archaeological field technique, which was later developed and improved in England and Wales by men such as Sir Mortimer Wheeler and Sir Cyril Fox.

20th-century developments. The 20th century has seen the extension of archaeology outside the areas of the Near East, the Mediterranean, and Europe, to other parts of the world. In the early '20s, excavations at Mohenjo-Daro and Harappā, in present Pakistan, revealed the existence of the prehistoric Indus civilization. In the late '20s, excavations at An-yang in eastern China established the existence of a prehistoric Chinese culture that could be identified with the Shang dynasty of early Chinese records.

Studies of the Stone Age

The Stone Age has been described and studied throughout the world; among the most sensational discoveries are those of L.S.B. Leakey, who found stone tools and skeletal remains of early man dating back 2,000,000 years in the Olduvai Gorge in Tanzania. Intensive work of great importance has brought to light early Neolithic sites at Jericho in Palestine; Hassuna, Iraq; Çatalhüyük, Turkey; and elsewhere in the Near East, establishing the origins of agriculture in that region.

Serious archaeological work began later in America than Europe, but as early as 1784 Thomas Jefferson had excavated mounds in Virginia and made careful stratigraphical observations. The 20th century has seen a great increase in archaeological knowledge about prehistoric America: two startling advances were the discovery of the origin of domesticated crops (including maize) in Central America and of the Olmec civilization of Mexico (1000–300 BC)—the oldest of the New World civilizations and probably the parent of all the others.

The enormous growth of archaeological work has meant the establishment of archaeology as an academic discipline; few important universities anywhere in the world are now without professors and departments of archaeology. There are now a very large number of scholarly journals in the field, as well as a considerable body of popularized books and journals that attempt to bridge the gap between professional and layman.

Fieldwork. *Preliminary work.* Some archaeologists call everything they do out-of-doors fieldwork, but others distinguish between fieldwork, in a narrower sense, and excavation. Fieldwork, in the narrow sense, consists of the discovery and recording of archaeological sites and their examination by methods other than the use of the spade and the trowel. Sites hitherto unknown are discovered by walking or motoring over the countryside: deliberate reconnaissance is an essential part of archaeological fieldwork.

In Europe, a study of old records and place-names may lead to the discovery of long-forgotten sites. The mapping of new and old sites is an essential part of archaeological survey. This process has been brought to a very high standard of perfection, both in the marking of archaeological sites on ordinary topographical maps and in the production of special period maps. The distribution map of artifacts, especially when studied against the background of the natural environment, is a key method of archaeological study.

Aerial photography in archaeology

The formerly earthbound archaeologist has been greatly helped by the development of aerial photography. The application of aerial photography to archaeological investigation began in a small way during World War I, as a side effect of military reconnaissance, and was given further impetus by World War II; the photographic intelligence departments of all the combatant nations were extensively staffed by archaeologists, who then carried their expertise and enthusiasm into the postwar years. The University of Cambridge now has its own department of air photography under J.K.S. St Joseph: using its own pilot and aircraft, it flies photographic missions over Ireland, Great Britain, Denmark, and The Netherlands. The number of new sites discovered each year by aerial photography is very large. Some of these are surface sites, especially partly destroyed sites that show up well in special conditions of light, as in early morning or late evening. But many are sites that could not be found on the ground and that show up in aerial photographs as variations in soil colour or in the density of crop.

Archaeological reconnaissance may be advanced from ordinary surface or aerial methods in a wide variety of ways. A very simple method is tapping the ground to sound for substructures and inequalities in the subsoil. Deep probes have made it possible to trace walls and ditches. The Lerici Foundation of Milan and Rome has had great success with this method since its development of the Nistri periscope, first used in 1957 in an Etruscan tomb in the cemetery of Monte Abbatone. The periscope is inserted into the burial chamber and can photograph the walls and contents of the whole tomb.

Other modern techniques that have been applied to archaeological prospecting employ electricity and magnetic fields (geophysical prospecting). A method of electrical prospecting had been developed in large-scale oil prospecting: this technique, based on the degree of electrical conductivity present in the soil, began to be used by archaeologists in the late 1940s and has since proved very useful. Magnetic methods of prospecting detect buried features by locating the magnetic disturbances they cause: these were introduced in 1957–58 and use such machines as the proton magnetometer, the proton gradiometer, and the fluxgate gradiometer. An American expedition discovered the site of Sybaris in Sicily by magnetic prospecting. Electromagnetic methods have been in use only since 1962; they employ developments of the concepts used in mine detectors. Instruments such as the pulsed-induction meter and the soil-conductivity meter detect magnetic soil anomalies, but only if the features are fairly shallow.

Excavation. Excavation is the surgical aspect of archaeology: it is surgery of the buried landscape and is carried out with all the skilled craftsmanship that has been built up in the last hundred years since Schliemann and Flinders Petrie. Excavations can be classified, from the point of view of their purpose, as planned, rescue, or accidental. Most important excavations are the result of a prepared plan—that is to say, their purpose is to locate buried evidence about an archaeological site. Many are project oriented: as, for example, when a scholar studying the life of the pre-Roman, Celtic-speaking Gauls of France may deliberately select a group of hill forts and excavate them,

as Sir Mortimer Wheeler did in northwestern France in the years before the outbreak of World War II. But many excavations, particularly in the heavily populated areas of central and northern Europe, are done not from choice but from necessity. Gravel digging, clearing the ground for airports, quarrying, road widening and building, the construction of houses, factories, and public buildings frequently threaten the destruction of sites known to contain archaeological remains. Emergency excavations then have to be mounted to rescue whatever knowledge of the past can be obtained before these remains are obliterated forever. Partial destruction of cities in western Europe by bombing during World War II allowed rescue excavations to take place before rebuilding. A temple of Mithras in the City of London, Viking settlements in Dublin and at Århus, Denmark, and the original 6th-century-BC Greek settlement of Massalia (Marseille) were discovered in this way. An extension of the runways at London Airport led to the discovery of a pre-Roman Celtic temple there.

Archaeology in urban areas

The role of chance in the discovery of archaeological sites and portable finds is considerable. Farmers have often unearthed archaeological finds while plowing their fields. The famous painted and engraved Upper Paleolithic cave of Lascaux in southern France was discovered by chance in 1940 when four French schoolboys decided to investigate a hole left by an uprooted tree. They widened a smaller shaft at the base of the hole and jumped through to find themselves in the middle of this remarkable pagan sanctuary. Similarly, the first cache of the Dead Sea Scrolls was discovered in 1947 by a Bedouin looking for a stray animal. These accidental finds often lead to important excavations. At Barnénès, in north Brittany, a contractor building a road got his stone from a neighbouring prehistoric cairn (burial mound) and, in so doing, discovered and partially destroyed a number of prehistoric burial chambers. The French archaeologist P.-R. Giot was able to halt these depredations and carry out scientific excavations that revealed Barnénès to be one of the most remarkable and interesting prehistoric burial mounds in western Europe.

All forms of archaeological excavation require great skill and careful preparation. Years of training in the field, first as an ordinary digger, then as a site supervisor, with spells of work as recorder, surveyor, and photographer, are required before anyone can organize and direct an excavation himself. Most museums, universities, and government archaeological departments organize training excavations. The very words dig and digging may give the impression to many that excavation is merely a matter of shifting away the soil and subsoil with a spade or shovel; the titles of such admirable and widely read books as Leonard Woolley's *Spadework* (1953) and *Digging Up the Past* (1930) and Geoffrey Bibby's *Testimony of the Spade* (1956) might appear to give credence to that view. Actually, much of the work of excavation is careful work with trowel, penknife, and brush. It is often the recovery of features that are almost indistinguishable from nonarchaeological aspects of the buried landscape: one example of this is the recovery of mud-brick walls in Mesopotamia; another is the tracing of collapsed walls of dry stone slabs in a cairn in stony country in the southwest Midlands of England. Sometimes it is the recovery of features of which only ghost traces remain, like the burnt-out bodies from the buried city of Pompeii, or the strings of a harp that were found among the furnishings of Mesopotamian tombs at Ur.

Excavation method

Because of the damage he may cause by inexperience and haste, the untrained amateur archaeologist often hinders the work of the professional. Amateur archaeology is forbidden in many countries by stringent antiquity laws. At the same time, it is certainly true that nonprofessionals have made important contributions in many areas of archaeology. Occasionally, an amateur does make an important discovery the further excavation of which can then be taken over by trained professionals. Such was the case at Sutton Hoo in Suffolk in 1939, when work begun by a competent amateur was taken over by a team of experts who were able to uncover a great Anglo-Saxon burial boat and its treasure, without doubt the most remarkable archaeological find ever made in Britain.

There are, of course, many different types of archaeological sites, and there is no one set of precepts and rules that will apply to excavation as a whole. Some sites, such as temples, forts, roads, villages, ancient cities, palaces, and industrial remains, are easily visible on the surface of the ground. Among the most obvious archaeological sites that have yielded spectacular results by excavation are the huge man-made mounds (tells) in the Near East, called in Arabic *tilāl,* and in Turkish *tepes* or *hüyüks*. They result from the accumulation of remains caused by centuries of human habitation on one spot. The sites of the ancient cities of Troy and Ur are examples. Another type consists of closed sites such as pyramids, chambered tombs, barrows (burial mounds), sealed caves, and rock shelters. In other cases there are no surface traces, and the outline of suspected structures is revealed only by aerial or geophysical reconnaissance as described above. Finally, there are sites in cliffs and gravel beds, where many Paleolithic finds have been made.

The wide range of techniques employed by the archaeologist vary in their application to different kinds of sites. The opening of the tomb chamber in an Egyptian pyramid is, for example, a very different operation from the excavation of a tell in Mesopotamia or a barrow grave in western Europe. Some sites are explored provisionally by sampling cuts known as sondages. Large sites are not usually dug out entirely, although a moderate-sized round barrow may be completely moved by excavation. Whatever the site and the extent of the excavation, one element of the technique is common to all digs, namely, the use of the greatest care in the actual surgery and in the recording of what is found by word, diagram, survey, and photography. To a certain extent all excavation is destruction, and the total excavation of a site subsequently engulfed by a housing estate or gravel digging is total destruction. This is why the archaeologist's field notes and his published report become primary archaeological documents. They are not themselves, strictly speaking, archaeological facts: they are the excavator's interpretation of what he saw, or thought he saw, but this is the nearest the discipline can ever get to archaeological facts as established by excavation. The really great excavators leave such a fine record of their digs that subsequent archaeologists can re-create and reinterpret what they saw and found. To delay publishing the results of an excavation within a reasonable time is a serious fault from the point of view of archaeological method. An excavation is not complete until the printed report is available to the world. Often the publication of the report takes as long as, or much longer than, the actual work in the field.

Publication of results

When a site like the Palace of Minos at Knossos or the city of Harappā in Pakistan has been excavated, and the excavations are over, the excavator and the antiquities service of the country concerned have to face the problem of what to do with the excavated structures. Should they be covered in again, or should they be preserved for posterity, and if preserved, what degree of conservation and restoration is permissible? This is the same kind of problem that arises in connection with the removal of antiquities from their homeland to foreign museums, and there is no generally accepted answer to it. These problems remain to beset archaeology: should Sir Arthur Evans have reconstructed the Palace of Minos at Knossos? Should the art treasures of ancient Greece and Egypt, now in western European museums, be returned? There is no simple, straightforward, overall answer to these difficult questions.

Underwater archaeology. Underwater archaeology is a branch of reconnaissance and excavation that has been developed only during the 20th century. It involves the same techniques of observation, discovery, and recording that are the basis of archaeology on land, but adapted to the special conditions of working underwater. It is obvious that no archaeologist working on submarine sites can get far unless he is trained as a diver. Helmeted sponge divers have made most of the important archaeological discoveries in the Mediterranean. The French scientist Jacques-Yves Cousteau developed the self-contained breathing ap-

paratus known as the scuba, of which the most commonly used type is the aqualung. Cousteau's work at Le Grand Congloué near Marseille was a pioneer underwater excavation, as was the work of the Americans Peter Throckmorton and George Bass off the coast of southern Turkey. In 1958 Throckmorton found a graveyard of ancient ships at Yassı Ada and then discovered the oldest shipwreck ever recorded, at Cape Gelidonya—a Bronze Age shipwreck of the 14th century BC. George Bass of the University of Pennsylvania worked on a Byzantine wreck at Yassı Ada from 1961 onward, developing the mapping of wrecks photogrammetrically with stereophotographs and using a two-man submarine, the "Asherah," launched in 1964. The "Asherah" was the first submarine ever built for archaeological investigation.

Interpretation. Excavation often seems to the general public the main and certainly the most glamorous aspect of archaeology; but fieldwork and excavation represent only a part of the archaeologist's work. The other part is the interpretation in cultural and historical contexts of the facts established—by chance, by fieldwork, and by digging—about the material remains of man's past. This task of interpretation has five main aspects.

Classification and analysis. The first concern is the accurate and exact description of all the artifacts concerned. Classification and description are essential to all archaeological work, and, as in botany and zoology, the first requirement is a good and objective taxonomy. Second, there is a need for interpretive analysis of the material from which artifacts were made. This is something that the archaeologist himself is rarely equipped to do; he has to rely on colleagues specializing in geology, petrology (analysis of rocks), and metallurgy. In the early 1920s, H.H. Thomas of the Geological Survey of Great Britain was able to show that stones used in the construction of Stonehenge (a prehistoric construction on Salisbury Plain in southern England) had come from the Prescelly Mountains of north Pembrokeshire; and he established as a fact of prehistory that over 4,000 years ago these large stones had been transported 200 miles from west Wales to Salisbury Plain. Detailed petrological analysis of the material of Neolithic polished stone axes have enabled archaeologists to establish the location of prehistoric ax factories and trade routes. It is also now possible, entirely on a petrological basis, to study the prehistoric distribution of obsidian (a volcanic glass used to make primitive tools).

Petrological analysis

In the third place, the archaeologist, having dealt with the material of his artifacts by classification and taxonomy, and with its physical nature by petrology and metallurgy, turns to the remaining information he can get from his colleagues in the natural sciences. These tell him the environmental conditions in which the people he is studying lived; he now sees his material remains not as isolated artifacts but in the context of their original environments.

Dating. Having analyzed his discoveries according to their form, material, and biological association, the archaeologist then comes to the all-important problem of dating. Many material remains of man's past have no dating problem: they may be, like coins, or most coins, self-dating, or they may be dated by man-made dates in written records. But the great and difficult part of the archaeologist's work is dating material remains that are not themselves dated. This can be done in one of three ways. Sometimes an object from another culture, the date of which is known (*e.g.,* in the case of pottery, by its style), is found at a previously undated site. Then, using the relative dating principle (see below) the archaeologist reasons that the material found with the imported object is contemporary with it. Conversely, an object from an undated culture may be found at a site whose date is known. Thus nonliterate communities can be dated by their contact with literate ones. This technique is known as cross dating; it was first developed by Sir Flinders Petrie when he dated Palestinian and early Greek (Aegean) sites by reference to Egyptian ones. Much of the prehistoric chronology of Europe in the Neolithic, Bronze, and Early Iron ages is based on cross dating with the ancient Near East.

Aside from cross dating, the archaeologist faced with material in a site having no literate chronological evidence of its own has two other ways of dating his material. The first is relative, the second absolute. Relative dating merely means the relation of the date of anything found to the date of other things found in its immediate neighbourhood. As has already been described, this method also plays a part in cross dating. Stratigraphy is the essence of relative dating. The archaeologist observes the accumulation of deposits in a gravel pit, a peat bog, in the construction of a barrow, or in accumulated settlements in a tell, and, like the geologists who introduced the principles of stratigraphy in the late 18th and early 19th centuries, he can see the succession of layers in the site and can then establish the chronology of different levels of layers relative to each other. In the excavation of a great tell like Ur or Troy the relative chronology of the various levels of occupation is the first thing to be established. Some archaeologists, even until quite recent times, have mistakenly supposed that depth below ground level is itself an indication of antiquity.

Relative dating

But even in properly observed and recorded stratigraphic levels there is often doubt, and the question arises: are all the artifacts and human remains found in the same level contemporary? Is it possible that there could have been later intrusions that have been difficult to distinguish in the field? The analysis of the fluorine content of bones has been very helpful here. Recognized as a valuable technique by French scientists in the 19th century, it was developed in England by K.P. Oakley in the 1950s. If bones in apparently the same geological or archaeological level have markedly different fluorine content, then it is clear that there must be interference—for example, by a later burial, or by deliberate planting of faked remains, as happened in the case of the Piltdown "Man" hoax in England.

Absolute man-made chronology based on king lists and records in Egypt and Mesopotamia goes back only 5,000 years. For a long time archaeologists searched for an absolute chronology that went beyond this and could turn their relative chronologies into absolute dates. Clay-varve counting seemed to provide the first answer to this need for a nonhuman absolute chronology. Called geochronology by Baron Gerard De Geer, its Swedish inventor, this method was based on counting the thin layers of clay left behind by the melting glaciers when the European Ice Age came to an end. This gave a chronology of about 18,000 years—three times as long as the man-made chronology based on Egyptian and Mesopotamian king lists. Thus, absolute dates could be established for artifacts from the Late Paleolithic Period, the whole of the Mesolithic Period, or Middle Stone Age, and much of the Early Neolithic Period.

Dendrochronology, the dating of trees by counting their growth rings, was first developed for archaeological purposes by A.E. Douglass in the United States. The application of this method to archaeology depends, obviously, on the use in antiquity of old datable trees in the construction of houses and buildings. It has been possible by dendrochronology to date prehistoric American sites as far back as the 3rd and 4th centuries BC.

The greatest revolution in prehistoric archaeology occurred in 1948, when Willard F. Libby, at the University of Chicago, developed the process of radioactive carbon dating. In this method, the activity of radioactive carbon (carbon-14) present in bones, wood, or ash found in archaeological sites is measured. Because the rate at which this activity decreases in time is known, the approximate age of the material can be determined by comparing it to carbon-14 activity in presently living organic matter. There have been problems and uncertainties about the application of the radioactive carbon method, but, although it is less than perfect, it has given archaeology a new and absolute chronology that goes back 40,000 years.

Radioactive carbon dating

Following the revolutionary discovery of radioactive carbon dating, other physical techniques of absolute dating were developed, among them potassium–argon dating and dating by thermoluminescence. Potassium–argon dating has made it possible to establish that the earliest remains of man and his artifacts in East Africa go back at least 2,000,000 years, and probably further.

Historical judgments. The last and most important task

of the archaeologist is to transmute his interpretation of the material remains he studies into historical judgments. When he is dealing with medieval and modern history he is often doing no more than adding to knowledge already available from documentary sources: but even so his contribution is often of great importance; for example, in relation to the growth and development of towns and the study of deserted medieval villages. When he is dealing with ancient history and prehistory, he is making a contribution of the greatest importance and often one that is more important than that of purely literary and epigraphical sources. For the prehistoric period, which now appears to stretch from 2,000,000 years ago to about 3000 BC, archaeological evidence is the only source of knowledge about human activities. But prehistoric remains have always been the most difficult to interpret, precisely because there are no written records to aid in the task. Now, with exact dating techniques at his disposal, the prehistorian is becoming more like the historical archaeologist and is concerned with the periodization and the historical contexts of his finds. (G.E.D.)

BIBLIOGRAPHY

Bibliography, the art or science of the description of books, has acquired special importance in the 20th century because of the need for effective organization of the records of human communication in the face of the enormous growth of publishing activity and the need, especially in undeveloped countries, for informed access to the world's scientific and technical information. It has been said that without bibliography, the records of civilization would be an uncharted chaos of miscellaneous contributions to knowledge, unorganized and inapplicable to human needs.

The word bibliography, in its literal sense, derived from the Greek *bibliographia* (2nd century AD), means the writing of books, and it was so defined in the 17th century; since the 18th century, it has been used to denote the systematic description and history of books. It is now commonly used in two widely divergent, though basically connected senses: (1) the listing of books, arranged according to some system (in this sense it is called enumerative, systematic, or descriptive bibliography); and (2) the study of books as material objects; *i.e.,* the study of the material of which books are made and the manner in which they are put together (in this sense commonly called critical bibliography). It is the function of bibliography to provide useful information for the student, in the one case supplying him with information about material for study, in the other helping him to establish the place of a book (or a piece of writing) in an author's production and its quality and authenticity as a text for study.

Kinds of bibliographies

Descriptive bibliography. The tasks of the compiler of a bibliography are (1) to find out what books on a particular subject exist; (2) to describe them item by item; and (3) to assemble the resulting entries into useful arrangements for reference and study. The need for lists of this kind arises as soon as the number of books in any subject is too great to be easily remembered. Among the earliest lists of books are those compiled by certain writers as guides to their own books—*e.g.,* the celebrated physician Galen compiled, in the 2nd century AD, a description in subject order of his own writings (*De propriis libris liber,* "A Book About My Own Books"); the Venerable Bede attached at the end of his *Ecclesiastical History of the English People,* in 731, an autobiographical note with a list of his writings; Erasmus also published a catalog of his own writings in narrative form in 1523. Early bibliographies, however, were not confined to such autobibliographies; as early as the last decade of the 4th century AD, the idea of attaching lists of works to lives of ecclesiastical writers was adopted by St. Jerome in his *De viris illustribus* ("Concerning Famous Men").

The invention of printing in the 1440s made possible the rapid multiplication of books, thereby increasing the need for guides to the resultant literature and also bringing about changes in the manner of compilation. The earliest substantial bibliography after the invention of printing was that of Johannes Tritheim, abbot of Sponheim in the diocese of Mainz, who in *Liber de scriptoribus ecclesiasticis* ("Book About Ecclesiastical Writers," 1494) included in chronological order, with an alphabetical index, about 1,000 ecclesiastical writers and their books. In 1545 a German-Swiss writer and naturalist, Conrad Gesner, who has been known as the father of bibliography, published his *Bibliotheca Universalis* (*Universal Bibliography*) of all Latin, Greek, and Hebrew writers, living and dead; this was followed three years later by a second volume, *Pandectarum sive Partitionum universalium libri XXI* ("Twenty-one Books of Encyclopedias or Universal Divisions [of Knowledge]"), in which the entries, arranged alphabetically in the earlier volume, are rearranged under 21 subject headings. Although Gesner was not the earliest descriptive bibliographer, his attempts at universality and classification earn him his fame. Gesner's idea of universality remained an ideal until recent times; in the face of the formidable problems of size, cost, and complexity, however, it has receded into the background of human aspirations. The Institut International de Bibliographie, founded in 1895 in Brussels by Henry Lafontaine and Paul Otlet with the object of creating a universal bibliography of books and articles in periodicals, arranged according to a specially designed system of classification, the Universal Decimal Classification (UDC), has assembled a card catalog of many millions of entries; but because of the immense bulk of the material and the growing cost of compilation, it is unlikely that it will ever reach the goal set by its founders. The hopes for universal bibliographies have been largely replaced by the published catalogs of such great comprehensive libraries as the British Museum (1961–67), the Bibliothèque Nationale (beginning in 1897), the Library of Congress, and, most promising of all, the United States National Union Catalogue, maintained in the Library of Congress and, in the early 1970's, in the course of being printed in an estimated 625 volumes. Investigations were (early 1970's) being made into the problems involved in programming and computerizing such catalogs, and if the complexities of programming multilanguage material can be solved, if the necessary financing can be found, and if world cooperation in the production and standardization of catalog entries can be assured, it seems possible that the ideals of the Brussels Institute may yet to some extent be realized.

Hopes for a universal bibliography

Gesner's achievement was remarkable; besides writing many other books, including a number of bibliographical works, his universal bibliography with its appendix (1555) describes and classifies some 15,000 works by about 3,000 writers and provides alphabetical and subject indexes. His contemporaries turned their attention to detailed bibliographies of writings on particular subjects or in particular languages or relating to particular countries or places. Thus, John Bale, bishop of Ossory, published the earliest national bibliography, *Illustrium maioris Britanniae scriptorum . . . summarium* ("Summary of the Writings of the Most Eminent Britons," 1548), which lists British writers chronologically and sets out their writings in detail. The vastly improved extended second edition of this work, *Scriptorum illustrium maioris Brytanniae . . . catalogus* (1557), acknowledges Bale's indebtedness to John Leland, Henry VIII's library keeper and antiquary, who had made an antiquarian tour through England in 1534–42 and had made a survey of the writings of British authors. Other national bibliographies followed, such as those of the Italian Antonio Francesco Doni, *La Libraria del Doni Fiorentino Nella quale sono scritti tutti gl'autori vulgari* ("Library of . . . All Secular Authors," 1550; 2nd ed. also 1550); *La seconda del Doni* (1551); the Dutch theologian Cornelius Loos, *Illustrium Germaniae scriptorum catalogus* (1582); the French bibliographer François Grudé de la Croix du Maine, *Premier Volume de la bibliothèque du Sieur de la Croix du Maine* ("The First Volume of the Library of . . . ," 1584). The catalogs of the German book fairs held in Frankfurt and Leipzig each spring and autumn, beginning in 1564 and continuing until 1749, while not bibliographies in the strict sense, were nevertheless widely used as foundation material by early German bibliographers.

Catalogs of the German book fairs

Types of descriptive bibliographies. Most countries now have national bibliographies, in a majority of cases pub-

lished officially by the national library and based on copies of national publications deposited in accordance with provisions of copyright acts. Notable exceptions are the United States and The Netherlands, whose national bibliographies are published commercially; East Germany, whose *Deutsche Nationalbibliographie* is published by the Deutsche Bücherei at Leipzig, an organization maintained since 1913 by German publishers; and the United Kingdom, whose *British National Bibliography* is published by a council representing libraries, publishers, and booksellers. These national bibliographies aim at, and attain, a high degree of completeness and promptitude. The *British National Bibliography* (beginning in 1950), for example, is published weekly and cumulated quarterly and annually; it is arranged in a classified order according to the decimal classification of the U.S. librarian Melvil Dewey, with an alphabetical index of authors, titles, and subjects. The *Bibliographie de la France* (beginning in 1811), published weekly, is arranged in a classified order with an annual index of authors, titles, and subjects. The East German *Deutsche Nationalbibliographie* (beginning in 1931) is published in two parts: (1) commercial publications, published weekly and (2) publications of societies and institutions, published semi-monthly. The West German *Deutsche Bibliographie* (beginning in 1947) is published weekly, with author and catchword index; this work has been published by computer since the beginning of 1968—the first such work to be so produced.

Bibliographies of books published in particular countries are of great value to students. Outstanding examples of such bibliographies are Charles Evans, *American Bibliography* (1903–34), covering the period 1639–1799; *A Short-Title Catalogue of Books Printed in England, Scotland and Ireland, and of English Books Printed Abroad, 1475–1640,* compiled by A.W. Pollard and G.R. Redgrave (1926; revised and enlarged 2nd ed., 1976), and its continuation by D.G. Wing, *Short-Title Catalogue . . . 1641–1700* (1945–51; revised 2nd ed., 2 vol., 1972–82). While not strictly bibliographies, the British Museum catalogs of its holdings of pre-1600 books from several European countries are similarly valuable because of the richness of the museum's collections.

Personal bibliographies may consist of no more than a simple list of an author's works; *e.g.,* those attached to articles in the *Dictionary of National Biography.* But they may be more elaborate—*e.g.,* F.W. Ebisch and L.L. Schücking, *A Shakespeare Bibliography* (1931; supplement 1937); Michael Sadleir, *Trollope: A Bibliography* (1928); Bertha Coolidge Slade, *Maria Edgeworth* (1937). Personal bibliographies are sometimes based on private collections—*e.g., A Stevenson Library* (1951–64) based on the collection of Edwin J. Beinecke; T.J. Wise's bibliographies, based on his own collections, of Tennyson, Swinburne, and others. A variant of personal bibliography consists of a narrative setting out an author's works with a comprehensive account of the circumstances surrounding the composition and publication of each work. An example is J.E. Norton, *A Bibliography of the Works of Edward Gibbon* (1940).

Subject bibliographies vary in size, scope, and method, according to the purpose they are designed to serve. The following are examples of bibliographies that aim at offering comprehensive guidance: A. and A. de Backer and A. Carayon, *Bibliothèque de la Compagnie de Jésus,* ("Library of the Society of Jesus," 9 vol.; new ed. 1890–1909); F.W. Bateson, *The Cambridge Bibliography of English Literature* (4 vol., 1940; new ed. 1970 ff.); *Bibliography of British History:* C. Read, *Tudor Period, 1485–1603* (1933), G. Davies, *Stuart Period, 1603–1714* (1928), S. Pargellis and D.J. Medley, *The Eighteenth Century, 1714–1789* (1951); P. Caron, *Manuel pratique pour l'étude de la Révolution française* ("Practical Manual for the Study of the French Revolution"; latest ed. 1947); F.C. Dahlmann and G. Waitz, *Quellenkunde der deutschen Geschichte,* ("Published Sources of German History"; 9th ed. 1931–32); W.W. Greg, *A Bibliography of the English Printed Drama to the Restoration* (4 vol., 1939–59); G. Lanson, *Manuel bibliographique de la littérature française moderne* ("Bibliographic Manual of Modern French Literature"; latest ed. 1947); F. Madan, *The Early Oxford Press* (3 vol., 1895–1931); L.-N. Malclès, *Les Sources du travail bibliographique* ("Sources of Bibliographic Work," 3 vol., 1950–52). Bibliographies of publications on particular subjects, frequently including articles on periodicals, are numerous and cover many branches of human knowledge and interests.

Bibliographies of bibliographies

The proliferation of bibliographies has led to the publication of bibliographical guides—bibliographies of bibliographies. Among these should be mentioned *Bibliotheca Bibliographica* (1866), compiled by Julius Petzholdt, librarian of the Royal Library at Dresden, Germany and *World Bibliography of Bibliographies* (5 vol.; 1965–66) by Theodore Besterman. Guides to current publications in a number of subject fields are published annually sometimes in narrative form—*e.g., The Year's Work in English-Studies* (1921 ff.) and *The Year's Work in Librarianship* (1929 ff.; later title *Five Years' Work in Librarianship,* 1958 ff.)—and sometimes in list form—*e.g., Internationale Bibliographie des Buch- und Bibliothekswesens* ("International Bibliography of Book and Library Systems," 1926–41).

Methods of compilation. The method of compiling a bibliography and the amount of detail included vary according to the author's purpose, his view of the importance of the subject, and his knowledge of it. The task of compiling a bibliography of any subject is a matter for a specialist in that subject. Having determined his objective and what material exists, the author must proceed to describe and arrange it as usefully for his purpose as he can. To ensure that the descriptions are accurate and consistent, they should be made from the works themselves. If, for any reason, this is impossible, the source of the information given should be stated.

For most bibliographies it is usually necessary to give only the author, short title, place of imprint, and date; but these should be transcribed accurately from the book or article. Notes may be added on the scope and quality of the text and the location of a copy. In an author bibliography a chronological arrangement is often adopted, enabling the user to follow the development of the author's work. In such cases, later editions are listed with the earliest edition but additional mention is made at the appropriate chronological point. Subject bibliographies may be arranged in a systematic order to bring out the features important for the author's purpose, or a recognized classification system, such as the Dewey Decimal Classification or Universal Decimal Classification, may be used.

Full descriptions are often given for early or rare books and for detailed author bibliographies. Elaborate rules have been evolved for compiling such descriptions, which make it possible for a skilled bibliographer to reconstruct from the text before him the makeup and appearance of a book. Such descriptions include information about the details of publication, number of copies printed, price, and binding. Semi-facsimile transcriptions are sometimes given, illustrating the various types of font used and the spacing of the title page. Modern methods of reproduction, however, make it possible to reproduce title pages photographically and thus obviate the use of such expedients, which in any case are not entirely satisfactory.

Use of reproductions

The use of reproductions, however, has certain disadvantages, such as high cost and difficulty of layout, and, more important, reproductions themselves may be misleading, especially when the quality of the printing and paper of the original are poor. In 17th-century printing, for example, it is sometimes difficult to distinguish a mark of punctuation from a flaw in the paper, and a reproduction may give a wrong impression. If facsimiles are used, they should be as carefully edited as the text and the source of the reproduction should be stated.

The advent of computers and the application of data-processing techniques to library procedures such as cataloging have great potential value in bibliography. If a catalog is maintained in machine-readable form, it can be printed out in a number of different arrangements, or a variety of sublistings can be produced, with a relatively small increase in costs over the cost of printing the original arrangement. This method depends on the planning of

requirements in advance and on the work of the programmer. It is thus possible for specialized lists, taken from the contents of a particular library, to be made available to a wide range of users. The best known and most highly sophisticated retrieval project of this kind is the Medlars project (Medical Literature Analysis and Retrieval System), which has as its main objective the publication of the *Index Medicus* (a monthly listing of current articles from about 2,300 biomedical journals throughout the world) but which also supplies printed listings on particular subjects in the field of medicine.

The full benefit of these methods will accrue when it is economically and administratively possible to contemplate the storage of the enormous amounts of information available in the large comprehensive libraries.

In 1968 the Library of Congress began the production and distribution of cataloging data in machine-readable form. The scope of this enterprise was limited initially to monographic works in the English language, but with the intention of expanding the coverage to include other languages. The creation of the machine record of material already in the library before 1969, which is necessary if a full system of machine-readable records is to be created, was being investigated both in the Library of Congress and elsewhere in other large comprehensive libraries. The problems likely to cause most difficulty are the labour involved, the great cost of the operation, and the magnitude of the data storage. Clearly such costly and complex operations have to be related to a realistic assessment of the benefits to be derived therefrom. Such assessments are in keeping with an interest, current in the early 1970s, in research library management. The rapid rise in library costs has emphasized the need for libraries, particularly large research libraries, to study whether and how they can apply accepted management techniques to the fuller exploitation of the bibliographical riches of their own institutions.

New approach to English literature

Critical bibliography. A valuable approach to the study of English literature, particularly, though not exclusively, of the 17th and 18th centuries, was developed about the turn of the 20th century out of a concern with Shakespearean texts in general and in particular with the nature and authority of the so-called quarto texts and their relation to the text printed in the First Folio. Editors had previously tended to assume that all the early texts of Shakespeare had equal authority and that it was perfectly proper to select readings which, in their view, improved the text of the author. The earlier editors were handicapped, in addition, by their ignorance of printing and book production in Shakespeare's time and by a lack of understanding of the nature of the texts they were dealing with. The change that took place stemmed from a variety of causes, of which the most important were probably two: (1) the close association in studies of English literature, especially drama, of three eminent scholars, A.W. Pollard, R.B. McKerrow, and W.W. Greg, all of them interested in bibliography; and (2) the remarkable progress made by Robert Proctor, a colleague of Pollard's in the British Museum, in the study of incunabula, the name given to books printed in the first 50 years after the invention of printing (*i.e.,* before 1501). The earliest printed books display the very individual characteristics of their printers in the type used, the typesetting, and the layout of the page, and it was found that, by studying these characteristics, useful deductions could be made about the place of individual books in a printer's total production and hence their approximate dates; in cases in which books contained no indication of the printer, it was often possible, using the same means, to assign a book to a particular printer. Important stages in the study of these earliest printed books, which had always been prized for their early date, their great aesthetic quality, and their rarity, were marked by a study of William Caxton, England's earliest printer, by William Blades, *The Life and Typography of William Caxton* (1861–63), in which typographical details were used to arrange Caxton's publications in chronological order; and by the work of a great antiquarian scholar and librarian, Henry Bradshaw, who made a special study of 15th-century books printed in the Netherlands. Bradshaw laid down an important principle: let the book speak for itself. This new method was still more widely applied by Robert Proctor, who succeeded in allotting the large collections of incunabula in the British Museum and the Bodleian Library at Oxford to places of origin, by countries and towns, and to printers. Proctor's work firmly established this method of study, and the definitive catalog of the incunabula in the British Museum, which he began, demonstrated the valuable results that can be obtained by the meticulous examination of all the features of a book—paper, type, makeup, ornamentation, sewing, binding, manuscript notes, and marks of ownership.

Examination of the physical features of a book

It was against this background and, applying these methods to the literary and dramatic work of English authors of the 16th and 17th centuries, that Greg, McKerrow, and Pollard developed what is now called critical, or analytical, bibliography, which has been succinctly described as the study of books as tangible objects. By using the evidence of the physical features of books, it was often found possible to arrive at conclusions about the place of a book in an author's production that are beyond the scope of literary judgment.

One of the earliest and most illuminating examples of the application of this method concerns the question of the priority between the two 1609 issues of Shakespeare's play *Troilus and Cressida.* These differ in the preliminary matter: the title of one describes the play as having been acted at the Globe Theatre; in the other, the title page makes no mention of a production of the play, and an epistle to the reader stresses this fact by stating that the piece had never been "clapper-clawed with the palmes of the vulger." Literary considerations would suggest that the latter issue was the earlier. An examination of the books themselves, however, shows that the title page of the former is in its original state, while the latter's original title page had been cut away and two leaves substituted, containing the new title page and the address to the reader, thus demonstrating, without question, the order of the two.

In his edition of Beaumont and Fletcher's *Elder Brother* (1905), Greg similarly provided an irrefutable demonstration of the correct chronological order of the two quarto editions of 1637. In Q1 an improperly adjusted quad or space lead had produced a mark above the line before the word "young" in Act I, scene 2, line 72, which was mistaken by the compositor printing Q2 for an apostrophe and printed as such. Hence, Q2 was shown to have been printed from Q1 and not from the author's manuscript.

The success of such demonstrations led to a much more detailed study of the physical features of books and the bibliographical deductions that can be made from them. The methods of critical bibliography have been substantially developed and have been widely applied to books of later periods, and it has been shown that, even in the sophisticated machine age, careful bibliographical examination has a significant role to play in determining the reliability of an author's text and the place of a book in an author's production.

Exposure of false first editions

The method was spectacularly displayed in the exposure of a number of alleged first editions of poems, essays, and other minor productions of well-known 19th-century authors, such as the "Reading" sonnets of Elizabeth Barrett Browning—*i.e.,* an edition of these sonnets, dated 1847, with Reading as the place of imprint. These books, more than 50 in number, were shown, in *An Enquiry into the Nature of Certain Nineteenth Century Pamphlets* (1934), by John Carter and Graham Pollard, to have been printed on kinds of paper, made from esparto grass or wood pulp, known not to have been in use at the dates shown on the title pages, and using printing types that were first cut much later.

It has been clearly demonstrated that machine-printed books of the 19th and 20th centuries produce textual problems rivalling those found in hand-printed books. While editorial procedures and the principles governing the interpretation of bibliographical evidence remain the same, new printing methods force bibliographers to devise new procedures to deal with the new problems presented to them.

A widely used guide for the bibliographical study of

English books up to the end of the 17th century is provided by R.B. McKerrow's *Introduction to Bibliography for Literary Students* (1927). The publications of the Bibliographical Society of London, especially its journal, *The Library* (1889 ff.), contain valuable information about bibliographical methods and their application. A useful account of the work of the society and a study by F.P. Wilson of the application of bibliographical techniques to the works of William Shakespeare, "Shakespeare and the New Bibliography," is contained in *The Bibliographical Society, 1892–1942: Studies in Retrospect,* (1945). Other bibliographical journals that may be usefully consulted are: *The Papers of the Bibliographical Society of America* (1904 ff.), the *Publications* of the *Oxford Bibliographical Society* (1922 ff.), and the *Transactions of the Cambridge Bibliographical Society* (1949 ff.). *Studies in Bibliography* (1948 ff.), published by the Bibliographical Society of the University of Virginia, is especially valuable. (F.C.F.)

CHRONOLOGY

Chronology, in the broadest sense, is a time scale, a method of ordering time. This article, however, deals mainly with systems of chronology used by different peoples in recording their history.

Scientific chronology, which seeks to place all happenings in the order in which they occurred and at correctly proportioned intervals on a fixed scale, is used in many disciplines and can be utilized to cover vast epochs. Astronomy, for example, measures the sequence of cosmic phenomena in thousands of millions of years; geology and paleontology, when tracing the evolution of the Earth and of life, use similar epochs of hundreds or thousands of millions of years. Geochronology reckons the more distant periods with which it deals on a similar scale; but it descends as far as human prehistoric and even historic times, and its shorter subdivisions consist only of thousands of years. Shortest of all are the chronological scales used in the recording of human events in a more or less systematic and permanent manner. These vary in scope, accuracy, and method according to the purpose, degree of sophistication, and skill of the peoples using them, as do the calendrical systems with which they are inextricably bound up. For further details see the article CALENDAR.

It is difficult to fix ancient historical chronologies in relation to scientific chronology. The terms of reference of ancient peoples were vague and inconsistent when judged by modern standards, and many of their inscriptions and writings have inevitably disappeared. The gaps in their records are increasingly filled in and their inconsistencies removed by the results of archaeological excavation. Guided by these findings, scholars can confirm, refute, or amend chronological reconstructions already tentatively made. Astronomical calculation and dating by radioactive-carbon content are also helpful in the work of fixing ancient chronologies.

Chinese. Chinese legendary history can be traced back to 2697 BC, the first year of Huang Ti (Chinese: Yellow Emperor), who was followed by many successors and by the three dynasties, the Hsia, the Shang, and the Chou. Recent archaeological findings, however, have established an authentic chronology beginning with the Shang dynasty, though the exact date of its end remains a controversial topic among experts. The so-called oracle-bone inscriptions of the last nine Shang kings (1324–1122 BC) record the number of months up to the 12th, with periodical additions of a 13th month, and regular religious services on the summer and winter solstice days, all of which indicates the adjustment of the length of the lunar year by means of calculations based on the solar year. Individual days in the inscriptions are named according to the designations in the sexagenary cycle formed by the combination of the 10 celestial stems and 12 terrestrial branches. Every set of 60 days is divided into six 10-day "weeks." Also recorded are numerous eclipses that can be used to verify the accuracy of the Shang chronology. In the oracular sentences of the last Shang king, Chou Hsin, the year of his reign is referred to as "the King's *n*th annual sacrifice."

From the beginning of the following (Chou) dynasty, the word year was etymologically identical with "harvest." Thus, "King X's *n*th harvest" meant the *n*th year of his reign. The lunar month was then divided into four quarters—Ch'u-chi, Tsai-sheng pa, Chi-sheng pa and Chi-szu pa—and the practice of using the 60 cyclical names for the days was continued. Thus, in the inscription on a Chou bronze vessel, a typical date would read: "In the King's *n*th harvest, in the *n*th quarter of the *n*th month, on the day X-*y*, etc."

The tradition of recording events by referring to the king's regnal year continued until 163 BC, when a new system, *nien-hao* ("reign-period title"), was introduced by Emperor Han Wen Ti of the Former Han dynasty (206 BC–AD 8). Thereafter, every emperor proclaimed a new *nien-hao* for his reign at the beginning of the year following his accession (sometimes an emperor redesignated his *nien-hao* on special occasions during his reign). A typical date in the *nien-hao* system might read, "the third year of the Wan-li reign period" (Wan-li san nien). In order to date any event in Chinese history, it is necessary to convert the year in the period of the designated *nien-hao* into the Western calendar.

The *nien-hao* system

During the Chou dynasty the civil year began with the new moon, which occurred before or on the day of the winter solstice. This "first month" of the Chou year (Chou *cheng*) was equivalent to the 11th month of the Hsia year (Hsia *cheng*) or to the 12th month of the Shang year. The first emperor, Shih Huang-ti, of the short-lived Ch'in dynasty (221–206 BC) made the year begin one month earlier—*i.e.,* with the lunation (the period of time between one new moon and the next) before the one in which the winter solstice occurred. The Ch'in year was continuously used until 104 BC, when Emperor Han Wu Ti promulgated the T'ai-ch'u calendar by reverting to the Hsia *cheng*—*i.e.,* by taking the third month of the Chou year, or the second lunation after the winter solstice, as the first month of the civil year. This lunar year (or Hsia *cheng*) was used till the last day of the Ch'ing, or Manchu, dynasty (1644–1911/12). When in 1911 the first republic was founded, the solar year was officially adopted, but successive governments kept the *nien-hao* tradition by referring any date to the number of years since the establishment of the republic—*e.g.,* 1948 was chronicled "the 37th year of the republic." In 1949, when the People's Republic of China was proclaimed, the old system was replaced by the Gregorian calendar.

(W.S.-cg.)

Japanese. The principal chronicles describing the origins of Japanese history are the *Nihon shoki* ("Chronicle of Japan") and the *Koji-ki* ("Record of Ancient Matters"). The *Nihon shoki* (compiled in AD 720) assembled information in a chronological order of days, months, and years starting several years before 660 BC, which was the year of the enthronement of the first Japanese emperor, who was posthumously named Jimmu. The *Koji-ki* (compiled in AD 712) related events under the reign of each emperor without a strict chronological order. Sometimes the *Koji-ki* gave the years of emperors' deaths and their ages at death. This information is different from that recorded in the *Nihon shoki.*

Native Japanese scholars since Fujiwara Teikan in the 18th century have realized that the *Nihon shoki* was historically inadequate and different from the *Koji-ki,* at least insofar as the chronological information is concerned. They have suggested that the foundation year of Japan was 600 years later than stated in the *Nihon shoki.* Naka Michiyo (late 19th century) argued with minute detail about the question of Japanese chronology. His ideas were supplemented by those of other Japanese scholars, who pointed out that: (1) the reigns of the earlier Japanese emperors as stated in the *Nihon shoki* are unnaturally long; (2) the date of the enthronement of the emperor Jimmu should be reconsidered; (3) a chronological gap exists between the *Nihon shoki* and contemporary Chinese and Korean chronicles. In comparison with Korean chronicles, they argued, the *Nihon shoki* has created an intentional expansion of chronology—*i.e.,* the entries about the empress Jingō and the emperor Ōjin can be identified with historical facts relating to the Korea of the 4th and 5th centuries and therefore must be placed 120 years later

Reforms of Naka Michiyo

than mentioned in the *Nihon shoki.* When comparing the *Nihon shoki* with Chinese chronicles, one finds the chronological gap somewhat reduced. The Chinese chronicles provide information about the tributes sent individually by five Japanese "kings" to Liu-Sung and Southern Ch'i during the 5th century. There are still questions about the identification of these kings, but it is generally accepted that the "king" written in Chinese character as Wu must be the Japanese emperor Yūryaku. By the late 5th century the gap between Japanese and Korean records, on the one hand, and Japanese and Chinese, on the other hand, disappears.

The intentional expansion of the chronology of the *Nihon shoki* was adopted by its compilers, who identified Queen Himiko (Pimihu) of Yamatai of the chronicle of Wei China with the Empress Jingō of Japanese legend.

The method of designating a year by the *kan-shi* (sexagenary cycle) appears to have begun about the reign of Emperor Yūryaku, when, as mentioned above, the gap between the continental and Japanese chronologies was bridged. The inscription on remarkable copper images of Buddha cast just after the period of Prince Shōtoku's regency (AD 593–621) bears a *nengō* (*nien-hao,* or reign-year title), although not a strictly authorized one. It was at this time that the Chinese luni-solar calendar system was adopted. The first official *nengō* was Taika, which was adopted by the imperial court in 645. Since 701, when the second title, Taihō, was adopted, the reign-year system has been continuously used in relation to the emperors' reigns up to the present day. In medieval times Japanese chronology underwent a remarkable evolution: (1) when the Imperial dynasty split into two courts (1336–92), two series of *nengō* began to be used; (2) during the Ashikaga period some private *nengō* again appeared; (3) some dates of the authorized "central" calendars did not correspond with those of locally compiled calendars. Moreover, military leaders would not accept some of the new *nengō.* Minamoto Yoritomo, for example, did not use the *nengō* that was adopted by the emperor Antoku and the Taira regime, and Ashikaga Mochiuji and Ashikaga Shigeuji did not use the official, respectively Eikyō and Kōshō, *nengō.*

In the Tokugawa period (1603–1867), gaps between central and provincial calendars disappeared, especially after the establishment of the Jōkyō calendar, the first native calendar compiled in Japan, instead of the Chinese-based one that was in use until this period. On January 1, 1873, Emperor Meiji adopted the Gregorian calendar in use in the West and at the same time adopted the "Japanese Era," with Emperor Jimmu as its founder, in addition to the *nengō* system. (Hi.Mo.)

Indian. Two kinds of chronological systems have been used in India by the Hindus from antiquity. The first requires the years to be reckoned from some historical event. The second starts the reckoning from the position of some heavenly body. The historical system, the more common in modern times, exists side-by-side with Muslim and international systems successively introduced.

Reckonings dated from a historical event. The inscriptions of the Buddhist king Aśoka (*c.* 265–238 BC) give the first epigraphical evidence of the mode of reckoning from a king's consecration (*abhiṣeka*). In these inscriptions (Middle Indian language in India or Greek and Aramaean in what is now Qandahār, Afghanistan) the dates are indicated by the number of complete years elapsed since the king's consecration. But the earlier existence of a reckoning of duration of reigns and dynasties is evidenced by the testimony of the Greek historian Megasthenes, who in 302 BC was the ambassador of Seleucus I Nicator, founder of the Seleucid Empire, to the court of Candragupta Maurya, Aśoka's grandfather. According to Megasthenes, the people of the Magadha kingdom, with its capital Pāṭaliputra (Patna), kept very long dynastic lists, preserved in the later Sanskrit Purāṇas (legends of the gods and heroes) and later Buddhist and Jain chronicles. They generally indicate, in years or parts of years, the duration of each reign.

Similar records of other periods and regions exist, and a relative chronology may be established. Unfortunately, it is not always possible to connect them with any absolute chronology, the precise dates of the reigns given being still unsettled. For example, in the Scythian period of the history of northern India, several inscriptions are dated from the beginning of the reign of Kaniṣka, the greatest king of the Asian (Kushān) invaders, but his dates are still uncertain (AD 78, 128–129, 144, etc., have been suggested for the beginning of a Kaniṣka era).

Other records give regnal years that can be linked with absolute chronology through other data—*e.g.,* those of several rulers of the Rāṣṭrakūṭa of the Deccan.

Hindu dynastic eras

The dynastic eras, founded by several rulers and kept up or adopted by others, are also numerous. The most important were the Licchavi era (AD 110), used in ancient Nepal; the Kalacuri era (AD 248), founded by the Abhūrī king Īśvarasena and first used in Gujarāt and Mahārāshtra and later (until the 13th century) in Madhya Pradesh and as far north as Uttar Pradesh; the Valabhī era (AD 318, employed in Saurāṣṭra) and the Gupta era (AD 320), used throughout the Gupta Empire and preserved in Nepal until the 13th century. Later came the era of the Thakuri dynasty of Nepal (AD 395), founded by Aṃśuvarman; the Harṣa era (AD 606), founded by Harṣa (Harṣavardhana), long preserved also in Nepal; the western Cālukya era (AD 1075), founded by Vikramāditya VI and fallen into disuse after 1162; the Lakṣmaṇa era (AD 1119), wrongly said to have been founded by the king Lakṣmaṇasena of Bengal and still used throughout Bengal in the 16th century and preserved until modern times in Mithilā; the Rājyabhīṣekasaka or Marāthā era (1674), founded by Śivājī but ephemeral.

Later, instead of the beginning of a reign or of a dynasty, the death of a religious founder was adopted as the starting point of an era. Among Buddhists the death of the Buddha and among the Jains the death of the Jina were taken as the beginning of eras. The Jain era (*vīrasaṃvat*) began in 528 BC. Several Buddhist sects (no longer existing in India) adopted different dates for the death (Nirvāṇa) of the Buddha. The Buddhist era prevailing in Ceylon and Buddhist Southeast Asia begins in 544 BC.

Historical events, now obscure, were the basis of the two most popular Indian eras: the Vikrama and the Śaka.

The Vikrama era (58 BC) is said in the Jain book *Kālakācāryakathā* to have been founded after a victory of King Vikramāditya over the Śaka. But some scholars credit the Scytho-Parthian ruler Azes with the foundation of this era. It is sometimes called the Mālava era because Vikramāditya ruled over the Mālava country, but it was not confined to this region, being widespread throughout India. The years reckoned in this era are generally indicated with the word *vikramasaṃvat,* or simply *saṃvat.* They are elapsed years. In the north the custom is to begin each year with Caitra (March–April) and each month with the full moon. But in the south and in Gujarāt the years begin with Kārttika (October–November) and the months with the new moon; in part of Gujarāt, the new moon of Āṣāḍha (June–July) is taken as the beginning of the year. To reduce Vikrama dates to dates AD, 57 must be subtracted from the former for dates before January 1 and 56 for dates after.

The Śaka era

The Śaka, or Salivāhana, era (AD 78), now used throughout India, is the most important of all. It has been used not only in many Indian inscriptions but also in ancient Sanskrit inscriptions in Indochina and Indonesia. The reformed calendar promulgated by the Indian government from 1957 is reckoned by this era. It is variously alleged to have been founded by King Kaniṣka or by the Hindu king Salivāhana or by the satrap Nahapāna. According to different practices, the reckoning used to refer to elapsed years in the north or current years in the south and was either solar or luni-solar. The luni-solar months begin with full moon in the north and with new moon in the south. To reduce Śaka dates (elapsed years) to dates AD, 78 must be added for a date within the period ending with the day equivalent to December 31 and 79 for a later date. For Śaka current years the numbers to be added are 77 and 78. The official Śaka year is the elapsed year, starting from the day following that of the vernal equinox. A normal year consists of 365 days, while the leap year has 366. The first month is Chaitra, with 30 days in a normal year and

31 in a leap year; the five following months have 31 days, the others 30.

A Nepalese era (AD 878) of obscure origin was commonly used in Nepal until modern times. The years were elapsed, starting from Kārttika, with months beginning at new moon. Another era, the use of which is limited to the Malabar Coast (Malayalam-speaking area) and to the Tirunelveli district of the Tamil-speaking area, is connected with the legend of the hero Paraśurāma, an avatar (incarnation) of the god Vishnu. It is called the Kollam era (AD 825). Its years are current and solar; they start when the Sun enters into the zodiacal sign of Virgo in north Malabar and when it enters into Leo in south Malabar. It is sometimes divided into cycles of 1,000 years reckoned from 1176 BC. Thus, AD 825 would have been the first year of the era's third millennium.

Eras based on astronomical speculation. During the period of elaboration of the classical Hindu astronomy, which was definitively expounded in the treatises called *siddhāntas* and by authors such as Āryabhaṭa I (born AD 476), Varāhamihira, Brahmagupta (7th century AD), etc., were developed the ancient Vedic notions on the cycle of years, embracing round numbers of solar and lunar years together. On the one hand, greater cycles were calculated in order to include the revolutions of planets, and the theory was elaborated of a general conjunction of heavenly bodies at 0° longitude after the completion of each cycle. On the other hand, cosmologists speculated as to the existence of several successive cycles constituting successive periods of evolution and involution of the universe. The period calculated as the basis of the chronology of the universe was the *mahāyuga,* consisting of 4,320,000 sidereal years. It was divided into four *yuga*s, or stages, on the hypothesis of an original "order" (*dharma*) established in the first stage, the Kṛta *yuga,* gradually decaying in the three others, the Tretā, Dvāpara, and Kali *yuga*s. The respective durations of these four *yuga*s were 1,728,000, 1,296,000, 864,000, and 432,000 years. According to the astronomer Āryabhaṭa, however, the duration of each of the four *yuga*s was the same—*i.e.,* 1,080,000 years. The basic figures in these calculations were derived from the Brahmanical reckoning of a year of 10,800 *muhūrta* (see CALENDAR: *The Hindu calendar*), together with combinations of other basic numbers, such as four phases, 27 *nakṣatra*s, etc. The movement of the equinoxes was at the same time interpreted not as a circular precession but as a libration (periodic oscillation) at the rate of 54 seconds of arc per year. It is in accordance with these principles that the calculation of the beginning of the Kali *yuga* was done in order to fix for this chronology a point starting at the beginning of the agreed world cycle. Such a beginning could not be observed, since it was purely theoretical, consisting of a general conjunction of planets at longitude 0°, the last point of the *nakṣatra* Revati (Pisces). It has been calculated as corresponding to February 18, 3102 BC (old style), 0 hour, and taken as the beginning of the Kali era. In this era, the years are mostly reckoned as elapsed and solar or luni-solar.

Cosmic cycles

In Hindu tradition the beginning of this era was connected with (1) events of the Mahābhārata war; (2) King Yudhiṣṭhira's accession to the throne; (3) 36 years later, King Parikṣit's consecration; and (4) Lord Krishna's death. Its years are still regularly given in Hindu almanacs.

An era resting upon a fictitious assumption of a complete 100-year revolution of the Ursa Major, the Great Bear (*saptarṣi*), around the northern pole was the Saptarṣi, or Laukika, era (3076 BC), formerly used in Kashmir and the Punjab. The alleged movement of this constellation has been used in Purāṇa compilations and even by astronomers for indicating the centuries.

Two chronological cycles were worked out on a basis of the planet Jupiter's revolutions, one corresponding to a single year of Jupiter consisting of 12 solar years and the other to five of Jupiter's years. The second, the *bṛhaspaticakra,* starts, according to different traditions, from AD 427 or from 3116 BC. Before AD 907 one year was periodically omitted in order to keep the cycle in concordance with the solar years. Since 907 the special names by which every year of the cycle is designated are simply given to present years of the almanac.

Side-by-side with Hindu and foreign eras adopted in India, several eras were created in the country under foreign influence, chiefly of the Mughal emperor Akbar: Bengali San (AD 593), Amli of Orissa and Vilayati (AD 592), Faṣlī (AD 590, 592 or 593 according to the district), and Sursan of Mahārāshtra (599). (J.L.A.F.)

Egyptian. At the end of the 4th millennium BC, when King Menes, the first king of a united Egypt, started his reign, the ancient Egyptians began to name each year by its main events, presumably to facilitate the dating of documents. These names were entered into an official register together with the height of the Nile during its annual inundation. Short notes at first, the year names developed into lengthy records of historical and religious events, especially of royal grants to the gods. These lists grew into annals, which were kept during the entire history of Egypt so that later kings could, after important events, consult the annals and ascertain whether a comparable occurrence had happened before. Unfortunately, these annals are lost. Only fragments from the 1st to the 5th dynasty (*c.* 3100–*c.* 2345 BC) are preserved, copied on stone. These fragments, however, are in such poor condition that they raise more chronological problems than they solve.

The Egyptian priests of the Ramesside period (*c.* 1300 BC) copied the names and reigns of the kings from Menes down to their time from the annals, omitting all references to events. Even this king list would have given a safe foundation of an Egyptian chronology, but the only extant copy, on a papyrus now kept at the Museo Egizio in Turin, has survived only in shreds, entire sections having been lost. Extracts from this king list, which name only the more important kings, are preserved in the temples of the kings Seti I and Ramses II at Abydos and on the wall of a private tomb at Ṣaqqārah (now in the Egyptian Museum), but they give little help in chronological matters.

The Turin Papyrus

When the Greeks began to rule Egypt after the conquest of Alexander the Great, King Ptolemy II Philadelphus, hoping to acquaint the new ruling class with the history of the conquered country, commissioned Manetho, an Egyptian priest from Sebennytus, to write a history of Egypt in the Greek language. As Manetho had access to the ancient annals, he added some of their entries to his list of kings and reigns, especially during the first dynasties. The more he progressed in time, the more he added semihistorical traditions and stories as they were composed by the Egyptian priests to discuss moral problems in the disguise of a historical "novel." There had been, undoubtedly, fewer historical facts in Manetho's history than one might expect. But Manetho's work, too, is lost except for some excerpts used by Sextus Julius Africanus and Eusebius in writing their chronicles. These, in turn, represented the material used in part by George Syncellus in the 8th century AD. During copying and recopying, Manetho's text clearly suffered many changes, unintentionally or by purpose. The figures of the reigns, especially of the older dynasties, for instance, were enlarged when some of the early Christian historians tried to equate King Menes with Adam. In addition, the excerpts were done carelessly. Therefore, Manetho's work, as handed down to us, is short of useless. Nevertheless, together with the fragments of the annals and of the king list of Turin, they create a framework of Egyptian chronology; so the division into dynasties was taken over from Manetho. But to achieve a continuous history of Egypt and to bridge the gaps left by the fragmentary state of the extant chronological material, scholars must turn to other means, particularly astronomical references found in dated texts. These are related principally to the rising of Sothis and to the new moon.

Theoretically, the Egyptian civil year began when the Dog Star, Sirius (Egyptian Sothis), could first be seen on the eastern horizon just before the rising of the Sun (*i.e.,* 19/20 of July). As the civil calendar of the ancient Egyptians consisted of 12 months (each of 30 days) and five odd days (called epagomenal days), the civil year was a quarter of a day too short in relation to the rising of Sothis, so that the new year advanced by one day every four years. New Year's Day and the rising of Sothis coincided again

The Sothic cycle

only after approximately 1,460 years, the so-called Sothic cycle. Dated documents mentioning the rising of Sothis can be translated into the present calendar by multiplying the number of days elapsed since the first day of the year by four and subtracting this sum from the date of the beginning of the particular Sothic cycle. The dates for the start of each Sothic cycle are fortunately known because the Roman historian Censorinus fixed the coincidence of New Year's Day and heliacal rising of Sothis in AD 139. Taking into account a slight difference between a Sothic year and a year of the fixed stars, the years 1322, 2782, and 4242 BC are taken as starting points of a Sothic cycle.

There are six ancient Egyptian documents extant giving Sothis dates, but only three of these are of value. The oldest is a letter from the town of Kahun warning a priest that the heliacal rising of Sothis will take place on the 16th day of the 8th month of year 7 of a king who, according to internal evidence, is Sesostris III of the 12th dynasty. This date corresponds to 1866 BC, according to the corrected Sothic cycle. The next date is given by a medical papyrus written at the beginning of the 18th dynasty, to which a calendar is added, possibly to ensure a correct conversion of dates used in the receipts to the actual timetable. Here it is said that the 9th day of the 11th month of year 9 of King Amenhotep I was the day of the heliacal rising of Sothis—*i.e.,* 1538 BC. This date, however, is only accurate provided that the astronomical observations were taken at the old residence of Memphis; if observed at Thebes in Upper Egypt, the residence of the 18th dynasty, the date must be lowered by 20 years—*i.e.,* 1518 BC. The third Sothis date shows that Sirius rose heliacally sometime during the reign of Thutmose III, which lasted for 54 years, on the 28th day of the 11th month; so year 1458 BC (point of observation at Memphis) or 1438 BC (point of observation at Thebes) must have belonged to the reign of this king. From these dates it is possible to calculate the absolute dates for the reigns of the 12th dynasty, as the durations of most of the reigns of the kings belonging to this dynasty are preserved on the king list of the Turin Papyrus. On the other hand, chronologists are able to compute the reigns of the kings of the 18th dynasty by utilizing the highest dates of their documents and the figures preserved by Manetho. Historians are also helped by the fact that the Egyptians sometimes identified a certain day as "exactly new moon"; they reckoned new moon from the morning after the last crescent of the waning moon had become invisible in the east just before sunrise. As there is a 25-year lunar cycle, such ancient Egyptian moon dates could be calculated with a fair amount of certainty but of course only if the ancient Egyptians themselves observed this celestial phenomenon accurately. There is some doubt, however, as it is shown by the attempts of very competent scholars to convert these moon dates. Sometimes even moon dates given by the same papyrus contradict themselves; in another case, the date given by a document had to be amended to achieve a reasonable result. These and other examples show that ancient Egyptian statements on celestial phenomena, especially on new moons, tend to be inaccurate because of faulty or inexact observations. Therefore, every date given for a fixed reign should be used with caution as the astronomical observation on which it is based may be inexact. Sometimes they are controlled by synchronism with Babylonian, Assyrian, or Hittite king lists or, later on, by the close interconnections between Greek and Egyptian history. Sometimes even biographical data are helpful. The statements found on small stelae inside the burial ground of the holy bulls of Memphis (Apis) register the dates of birth, enthronement, and death of these animals accurately. But the more time recedes, the more the chronology of the Egyptian history becomes uncertain, even when astronomical data are available. Up till now even carbon-14 data are of no great help, as uncertainties are mostly not greater than the standard deviations to be expected in a carbon-14 calculation.

Nevertheless, Egyptologists believe themselves to be on fairly firm ground when dating the beginning of the Ancient Kingdom (1st and 2nd dynasty) about 3090 BC, the beginning of the 11th dynasty at 2133 BC, and of the Middle Kingdom (12th dynasty) at 1991 BC. The New Kingdom started at 1567 or 1552 BC, depending on a choice for the first year of Ramses II of either 1290 BC or 1304 BC—one lunar cycle earlier. The following centuries still pose many chronological questions down to 664 BC, when Greek historiography took over. (W.H.)

Babylonian and Assyrian. *Mesopotamian chronology, 747 to 539 BC.* The source from which the exploration of Mesopotamian chronology started is a text called Ptolemy's Canon. This king list covers a period of about 1,000 years, beginning with the kings of Babylon after the accession of Nabonassar in 747 BC. The text itself belongs to the period of the Roman Empire and was written by a Greek astronomer resident in Egypt. Proof of the fundamental correctness of Ptolemy's Canon has come from the ancient cuneiform tablets excavated in Mesopotamia, including some that refer to astronomical events, chiefly eclipses of the Moon. Thus, by the time excavations began, a fairly detailed picture of Babylonian chronology was already available for the period after 747 BC. Ptolemy's Canon covers the Persian and Seleucid periods of Mesopotamian history, but this section will deal only with the period up to the Persian conquest (539 BC).

The chief problem in the early years of Assyriology was to reconstruct a sequence for Assyria for the period after 747 BC. This was done chiefly by means of *limmu,* or eponym, lists, several of which were found by early excavators. These texts are lists of officials who held the office of *limmu* for one year only and whom historians also call by the Greek name of eponym. Annals of the Assyrian kings were being found at the same time as eponym lists, and a number of these annals, or the campaigns mentioned in them, were dated by eponyms who figured in the eponym lists. Moreover, some of the Assyrian kings in the annals were also kings of Babylonia and as such were included in Ptolemy's Canon.

The *limmu* lists

Good progress was therefore being made when, soon after 1880, two chronological texts of outstanding importance were discovered. One of these, now known as King List A, is damaged in parts, but the end of it, which is well preserved, coincides with the first part of Ptolemy's Canon down to 626 BC. The other text, *The Babylonian Chronicle,* also coincides with the beginning of the canon, though it breaks off earlier than King List A. With the publication of these texts, the first phase in the reconstruction of Mesopotamian chronology was over. For the period after 747 BC, there remained only one serious lacuna—*i.e.,* the lack of the eponym sequence for the last 40 years or so of Assyrian history. This had not been established by the early 1970s.

Assyrian chronology before 747 BC. German excavations at Ashur, ancient capital of Assyria, yielded further eponym lists. By World War I the full sequence of eponyms was known from about 900 to 650 BC. A further fragmentary list carried the record back to about 1100 BC, and on this basis Assyrian chronology was reconstructed, with little error, back to the first full regnal year of Tiglathpileser I in 1115 BC. Without another eponym list, a king list was needed for substantial further progress. King lists found at Ashur proved disappointing. Those fairly well preserved did not include figures for the reigns, and those with figures were very badly damaged.

In 1933, however, an expedition from the University of Chicago discovered at Khorsabad, site of ancient Dur Sharrukin, an Assyrian king list going back to about 1700 BC. But for the period before 1700 BC the list is damaged and otherwise deficient, and Assyrian chronology prior to this date is still far from clear.

Before 747 BC it was the custom of the Assyrian kings to hold eponym office in their first or second regnal year. Thus, in an eponym list, the number of names between the names of two successive kings usually equals the number of years in the reign of the first of the two kings. It would have been easy to compile a king list from an eponym list, and there is evidence that this Assyrian king list was compiled from an eponym list probably in the middle of the 11th century BC. As an eponym list is a reliable chronological source, since omission of a name entails an error of only one year, the king list, if based

on one, will have preserved much of the structure of older eponym lists now lost. (Except for one fragment, no known eponym list goes back further than the beginning of the 11th century BC.)

Babylonian chronology before 747 BC. In the long interval between the fall of the last Sumerian dynasty *c.* 2000 BC and 747 BC there are two substantial gaps in chronology, each about two centuries long. The earlier gap is in the 2nd millennium, from approximately 1600–1400 BC, the later gap in the 1st millennium, from *c.* 943–747 BC. During these gaps the names of most of the kings are known, as well as the order, but usually not the length of their reigns.

A means of checking the reliability of the Babylonian king list is provided by the chronicles, annals, and other historical texts that show that a given Assyrian king was contemporaneous with a given Babylonian king. There are no fewer than 15 such synchronisms between 1350 and 1050 BC, and, when the Babylonian and Assyrian king lists are compared, they all fit in easily. Only one of them, however, provides a close approximate date in Babylonian chronology. This synchronism shows that the two-year reign of the Assyrian king Ashared-apil-Ekur (*c.* 1076–*c.* 1075 BC) is entirely comprised within the 13-year reign of the Babylonian king Marduk-shapik-zeri. The Assyrian's dates are probably correct to within one year. Thus, if Marduk-shapik-zeri is dated so that equal proportions of his reign fall before and after that of Ashared-apil-Ekur, a date is obtained for the former that should not be in error more than six years. This synchronism constitutes a key to the structure of Babylonian chronology by providing the base date for all the reigns in the interval *c.* 1400–943 BC for which the Babylonian king list gives figures. All the dates thus obtained are subject to the six-year margin of error.

These synchronisms between Assyrian and Babylonian kings continue throughout the period that corresponds to the second gap in the Babylonian king list—from *c.* 943–747 BC. Since the Assyrian chronology in that period is firmly established, these synchronisms provide a useful framework for the structure of Babylonian chronology in that period.

The gap in the 2nd millennium BC, however, is not as easy to fill. The fact that the magnitude of the gap is uncertain constitutes the main problem in the chronology of the 2nd millennium BC and also affects the chronology of the preceding Sumerian period. The problem is not yet solved. Observations of the planet Venus made during the reign of King Ammisaduqa, less than 50 years before the end of the 1st dynasty of Babylon, permit only certain possible dates for his reign. Translated into dates for the end of the dynasty, the three most likely possibilities are 1651, 1595, and 1587 BC. The evidence is not yet conclusive and leaves uncertain what choice should be made among the three. The chronology adopted here is based on the second of these dates for the end of the 1st Babylonian dynasty—*i.e.,* 1595 BC.

Prior to this gap in the 2nd millennium BC, there is a period of five centuries with a well-established chronological structure. All the kings in the major city-states are known, as well as their sequence and the length of their reigns. Which sets of dates should be assigned to these reigns, however, depends on the date adopted for the 1st dynasty of Babylon. This period of five centuries extends from the beginning of the 3rd dynasty of Ur to the end of the 1st dynasty of Babylon—*i.e.,* on the chronology adopted here, 2113–1595 BC. During this period the Babylonians dated their history not by regnal years but by the names of the years. Each year had an individual name, usually from an important event that had taken place in the preceding year. The lists of these names, called year lists or date lists, constitute as reliable a source in Babylonian chronology as the eponym lists do in Assyrian chronology. One of the events which almost invariably gave a name to the following year was the accession of a new king. Hence, the first full regnal year of a king was called "the year (after) NN became king." In Assyria the number of personal names in an eponym list between the names of two successive kings normally equalled the number of years in the reign of the first king, and, similarly, in Babylonia the number of year names between two year names of the above kind nearly always equalled the number of years in the reign of the first king. Just as in Assyria, the eponym lists are almost certainly the source of the king lists, so in Babylonia the king lists are based on the year lists. Several of these king lists, compiled at a time when the year lists were still in use, survive. One gives the 3rd dynasty of Ur and the dynasty of Isin; another gives the dynasty of Larsa. Both may be school texts.

abylonian ear lists

The 3rd dynasty of Ur and the dynasty of Isin also figure in the Sumerian king list, which reaches far back into the Sumerian period. The original version probably ended before the 3rd dynasty of Ur, but later scribes brought it up to date by adding that dynasty as well as the dynasty of Isin.

(M.B.R.)

Jewish. The era at present in vogue among the Jews, counted from the creation of the world (*anno mundi;* abbreviated to AM), came into popular use about the 9th century AD. Traceable in dates recorded much earlier, this era has five styles conventionally indicated by Hebrew letters used as numerals and combined into mnemonics, which state the times of occurrence of the epochal mean conjunctions of *moladim* (see CALENDAR: *Middle Eastern calendar systems: The Jewish calendar*) or the orders of intercalation in the 19-year cycle or both. The respective epochs of these styles fall in the years 3762–3758 BC, inclusive. By about the 12th century AD the second of the mentioned styles, that which is in use at present, superseded the other styles of the era *anno mundi.*

The styles of this era arise from variations in the conventional rabbinical computation of the era of the creation. This computation, like hundreds of other calculations even more variable and no less arbitrary, is founded on synchronisms of chronological elements expressed in the terms of biblical and early postbiblical Jewish eras.

The biblical era *anno mundi* underlies the dating of events (mainly in the book of Genesis) prior to the Exodus from Egypt. This period of biblical chronology abounds in intractable problems caused by discrepancies between the Jewish and Samaritan Hebrew texts and the Greek version known as the Septuagint, by apparent inconsistencies in some of the synchronisms, and by uncertainties about the method of reckoning.

Discrepancies in biblical texts

During the period from the Exodus to the founding of Solomon's Temple, the only continuous biblical era (chiefly in the remaining books of the Pentateuch) is the era of the Exodus. With regard to a crucial date expressed in this era—"In the four hundred and eightieth year after the people of Israel came out of the land of Egypt, in the fourth year of Solomon's reign over Israel, in the month of Ziv, which is the second month, he began to build the house of the Lord" (I Kings 6:1)—there is again a discrepancy between the Hebrew text and the Septuagint. Other problems to be met with during this period are due to the obscurity of chronological data in the book of Judges and in I and II Samuel.

Dating by the kings of Israel

During the following period, the Bible uses the eras of the regnal years of monarchs (the kings of Judah, Israel, and Babylon) and of the Babylonian Exile. This period of biblical chronology likewise poses numerous problems, also the result of apparent inconsistencies of the synchronisms—*e.g.,* in the period from the accession of Rehoboam of Judah and of Jeroboam of Israel to the fall of Samaria "in the sixth year of Hezekiah [of Judah], which was the ninth year of Hoshea king of Israel" (II Kings 18:10) the reigns of the southern kingdom exceed those of the northern kingdom by 25 years.

The biblical data might be easier to harmonize if the occurrence of coregencies were assumed. Yet, as an ever-variable factor, these evidently would not lead to the determination of the true chronology of this period. Scholars therefore seek additional information from sources outside the Bible—*e.g.,* inscriptions on Assyrian monuments, which are dated by the so-called eponym lists. Substantial use also has been made of the data in the king list known as Ptolemy's Canon (compiled in the 2nd Christian century) commencing in 747 BC with the reigns of the Babylonian

kings (see above *Babylonian and Assyrian*). Scholars differ widely, however, in their interpretation of details, and numerous chronological problems remain unsolved. Only a few dates in this period can be fixed with any degree of confidence.

Reckoning by Persian kings

After the Babylonian Exile, as evidenced by the data in the Bible and the Aswān papyri, the Jews reckoned by the years of the Persian kings. The chronological problems of this period are caused by the apparent disorder in the sequence of events related in the biblical books of Ezra and Nehemiah and by the difficulty of identifying some of the Persian kings in question. For example, the King Artaxerxes of these books may stand for Artaxerxes I Longimanus (465–425 BC), for Artaxerxes II Mnemon (404–359/358 BC), or in the case of Ezra at any rate, for Artaxerxes III Ochus (359/358–338/337 BC).

From the Grecian period onward, Jews used the Seleucid era (especially in dating deeds; hence its name Minyan Shetarot, or "Era of Contracts"). In vogue in the East until the 16th century, this was the only popular Jewish era of antiquity to survive. The others soon became extinct. These included, among others, national eras dating (1) from the accession of the Hasmonean princes (*e.g.,* Simon the Hasmonean in 143/142 BC) and (2) from the anti-Roman risings ("era of the Redemption of Zion") in the years 66 and 131 of the Common (Christian) Era. Dates have also been reckoned from the destruction of the Second Temple (*le-ḥurban ha-Bayit*). The various styles of the latter, as also of the Seleucid era and of the era *anno mundi,* have often led to erroneous conversions of dates. The respective general styles of these eras correlate as follows: 3830 AM = year 381 of the Seleucid era = year 1 of the Era of the Destruction = year 69/70 of the Common (Christian) Era.

Earlier Jewish chronologies

The earliest Jewish chronologies have not survived. Of the work of the Alexandrian Jew Demetrius (3rd century BC), which deduced Jewish historical dates from the Scriptures, only a few fragments are extant. In the *Book of Jubilees,* events from the creation to the Exodus are dated in jubilee and sabbatical cycles of 49 and 7 years, respectively. Scholars differ as to the date and origin of this book. The era of the creation therein is unlikely to have been other than hypothetical.

The earliest and most important of all Jewish chronologies extant is the *Seder ʿolam rabbaʾ* ("Order of the World"), transmitted, according to Talmudic tradition, by Rabbi Yosi ben Halafta in the 2nd century AD. The author was possibly the first to use the rabbinic Era of the Creation. His chronology extends from the creation to Bar Kokhba in the days of the Roman emperor Hadrian (2nd century AD); but the period from Nehemiah to Bar Kokhba (*i.e.,* from Artaxerxes I or II to Hadrian) is compressed into one single chapter. The Persian phase shrinks to a mere 54 years. The smaller work *Seder ʿOlam zuṭaʾ* completes the *Rabbaʾ*. It aims to show the Babylonian exilarchs as lineal descendants of David.

Megillat taʿanit ("Scroll of Fasting"), although recording only the days and months of the year without the dates of the years, is nevertheless an important source for Jewish chronology. It lists events on 35 days of the year that have been identified with events in five chronological periods: (1) pre-Hasmonean, (2) Hasmonean, (3) Roman (up to AD 65), (4) the war against Rome (65–66), and (5) miscellaneous. The authors, or rather the last revisers, are identified with Zealots guided by Hananiah ben Hezekiah ben Gurion and his son Eliezer. (E.J.Wi.)

Greek. As the cities of ancient Greece progressed to their classical maturity, the need arose among them for a chronological system on a universally understood basis.

Dating by local records

In the archaic period, genealogies of local monarchs or aristocrats sufficed for the historical tradition of a given area, and events were associated with the lifetimes of well-known ancestors or "heroes." The *synoikismos* (founding of the united city) of Athens took place "in the time of Theseus"; the Spartan ephorate (chief magistracy) was established "in the reign of King Theopompus." When the city-states adopted annual magistracies, the years were designated by the eponymous officials—"in the archonship of Glaucippus" or "when Pleistolas was ephor." This was the local usage throughout classical and Hellenistic Greece, the title of the magistrate varying in different cities. Sometimes tenure of a priesthood provided the chronological basis, as at Argos, where years were dated as the *n*th of the (named) priestess of Hera. The correctness of the series was a matter first of memory and later of careful record. The list of annual archons at Athens was known back to 683 BC (in modern terms). Lists of dynasties also amounted to recorded folk memory, and in all genealogical reckoning there is a point, for modern critics, at which acceptable tradition shades into myth. Corruption of the records was introduced through error or political design, and traditions often conflicted.

Chronology became subject to systematization when cities felt a national need for accurate clarification of their past. In literature the growth of historiography initiated a search for a method of dating that could be universally applied and acknowledged. In the 5th and 4th centuries, local historians used local magistracies as their framework; research was devoted to rationalization of conflicting traditions and production of definitive lists. Charon of Lampsacus, perhaps in the early 5th century, compiled a record of Spartan magistrates; Hellanicus of Lesbos, author of the earliest history of Athens, wrote on the priestesses of Argos. Lists of victors in the great Olympic games were valid for all Greece, pointing the way to the widely accepted reckoning by Olympiads (see below). The Athenian Philochorus was the latest (early 3rd century BC) of compilers of Olympionikai.

The 5th-century historian Herodotus relied for his chronology principally upon the reckoning by generations used by his informants, conventionally accepted as showing three generations to a century. In some cases a 40-year, or other, reckoning was used, and varying traditions sometimes produced difficulty of synchronism. Thucydides, writing "contemporary" history, recognized the chronological problems involved. He dated the beginning of the Peloponnesian War by the Athenian, Spartan, and Argive systems and thenceforward marked the passage of time by seasonal indications. Synchronization was not helped by the fact that the official year began at different times in different cities. In later historical writing the impossibility of accurately coordinating the Athenian and Roman years resulted in serious chronological difficulties.

Dating by the Olympiads

The system of dating by Athenian archons came to be recognized outside Attica as of wider value, but, in the Hellenistic period, Alexandrian scholarship, represented especially by Eratosthenes of Cyrene, the "father of chronology," was instrumental in promoting the use of the Olympiads as an acceptable system, reckoning a four-year period from each celebration of the Olympic Games. Timaeus of Tauromenium (*c.* 356–260 BC) was the first historian to employ it, but it was little used outside historical writing. Aristotle had been concerned to identify the generation of the first Olympiad, accepted as 776 BC on modern reckoning. For convenience, the beginning of the Olympic year was equated with the summer solstice, when the Athenian year also began. This makes it generally necessary for a Greek year to receive a double date in modern terms (*e.g.,* the death of the philosopher Epicurus in 271/270 BC). Eratosthenes' system produced tables of dates, from which, for example, the fall of Troy could be dated to 1184/83 BC. The "Parian Marble" of 264/263 BC is an inscribed record of events from the time of Cecrops, first king of Athens, reckoning years between the date of the inscription, fixed by the Athenian archon, and each event concerned. Some cities inscribed lists of their eponymous magistrates; the Athenians were the first to do so *c.* 425 BC. A list from Sicilian Tauromenium originally spanned some 300 years. The regnal years of the Hellenistic monarchs or the count from a fixed event (a city foundation or refoundation) also provided acceptable chronological reckoning often useful for more than contemporary or local purposes.

The use of these chronological possibilities is best seen in historians using the annalistic method, of whom Diodorus Siculus is most notable. In the Christian period, Eusebius, followed by St. Jerome, began the work of reconciling all these indications to the Judaic tradition and produced

the foundation of chronology in terms of the Julian calendar upon which modern historians have constructed their framework.

For modern scholarship the problem, in E.J. Bickerman's words, is "how we know Caesar was assassinated on March 15, 44 BC." Before 480 BC, no date can be precise in terms of the Julian calendar unless confirmed by astronomical phenomena. Archaic chronology relies upon the typology of Corinthian pottery in relation to the foundation dates for Greek colonies in Sicily implicit in Thucydides, book vi. Julian dates given for this period (*e.g.,* for the tyranny of Peisistratus in Athens) stem from a complex combination of ancient chronographic tradition with modern archaeology, acceptable only with appropriate reserve. Literary tradition gives the succession of Athenian archons from 480 to 294 BC. The regnal, era, and Olympiad years also provide dates within a 12-month period. Closer dating is seldom possible unless the sources give precise information in calendric terms, as occasionally in literature and regularly in Athenian and Egyptian public documents. Even these are not translatable into Julian months and days unless coordinated with knowledge of contemporary solar or lunar phenomena and of possible official interference with the calendar.

(A.G.W.)

Roman. The establishment of a sound chronology for Roman history, as for Greek, depends on the assessment of the evidence available, which falls into two categories—literary and archaeological.

Literary evidence. Although by the late 3rd century BC the Greek mathematician Eratosthenes was working on the systematization of chronography and a series of learned historians had used the documentary method—*e.g.,* for Roman history, Timaeus of Tauromenium, to whom are probably due many of the synchronizations of Roman history with the Greek Olympiads—unfortunately this tradition of documentation and concern for chronology did not immediately pass over into Roman historiography. According to Cicero in *De oratore,* the earliest Roman historians did no more than "compile yearbooks"—for example, Fabius Pictor in the late 3rd century BC, Lucius Calpurnius Piso in the 2nd, and the so-called Sullan annalists in the 1st. Of these authors it is possible to judge only at second hand, and only those of the 1st century were much used directly by the historians whose work survives in any quantity, notably Livy, Dionysius of Halicarnassus, and Diodorus Siculus. In these authors, as in other 1st-century historians such as Sallust, there is little concept of documentation or research other than comparison of literary sources; for none was chronology a direct concern, and in many cases dramatic effectiveness took priority over fidelity to truth. Apart from the Greek Polybius, who treated the rise of Roman power in the Mediterranean from 264 to 146 BC, it was not until Cicero's time that the conception of historical scholarship developed in Rome. Cicero's friend Atticus not only was concerned to draw up a chronological table in his *Liber annalis* but had undertaken research to that end, and the great scholar Marcus Terentius Varro and a little later the learned Marcus Verrius Flaccus produced a vast body of erudite work, nearly all lost. To this source must probably be ascribed the *Fasti Capitolini,* a list of magistrates from the earliest republic to the contemporary period, set up near the *regia* (the office and archive of the *pontifices,* or high priests), perhaps on the adjacent Arch of Augustus, at the end of the 1st century BC. This work, since it is based on inscriptions, is sometimes given precedence over literary evidence, but, since it is a compilation, it is still subject to serious error.

Magistrate lists

Sources used by Roman historians. The traditionally early extant bodies of law, such as the Twelve Tables from the early republic, were of little chronological value, and juristic *commentarii* were liable to mislead through their zeal for precedent, while Cicero, in spite of Polybius' claim to have inspected early treaties preserved in the Capitol definitely states that there were no public records of early laws. A source frequently referred to is the *Annales maximi,* a collection made about 130 BC of the annual notices displayed on a white board by the *pontifices* and containing notes of food prices, eclipses, etc. Dionysius of Halicarnassus implied that they gave a date for the foundation of the city but was reluctant to accept their authority; and one of the eclipses is referred to by Cicero as being mentioned also by Ennius, but unfortunately the number of the year "from the foundation of the city" is corrupt in the text. Although it is possible to calculate the dates of eclipses astronomically in terms of the modern era, it is difficult to link these to Roman chronology because of the uncertainty of the figures and because of the confused state of the Roman calendar before the Julian reform (see CALENDAR: *Early calendar systems: The early Roman calendar*). Another difficulty is that the early records may have been burned in 390 BC when Celtic tribes sacked the city; also they would probably have been largely unintelligible if authentic.

Livy quoted the 1st-century annalist Gaius Licinius Macer as having found in the temple of Juno Moneta "linen rolls" giving lists of magistrates; but he also said that Macer and Quintus Aelius Tubero both cited the rolls for the consuls of 434 but gave different names. In any case, it is unlikely that the list could have been older than the temple, which dates from 344 BC. It is clear that the chief sources for the lists were the pedigrees of prominent Roman families, such as the Claudii Marcelli, Fabii, and Aemilii, drawn up by Atticus; but Cicero and Livy agree that tendentious falsifications had in many cases corrupted the records, and other suspicious facts are the appearance of obviously later or invented *cognomina,* or third names, and of plebeian gentile names for the earliest period, when only patricians bore them. Many scholars, however, accept the general authenticity of the lists—one reason being the appearance in them of extinct patrician families—but prefer Livy's version to that of the Capitoline lists, which show signs of late revision, often give names in incorrect order, and contain other anomalies.

Fixing the terminal date

The question, therefore, remains whether Roman chronography was dependent on the lists of magistrates or whether these were adapted to fit other known datings. The apparent advantages of the existence of a terminal date, the "foundation of the city," is illusory for Roman chronology, since it depended on back reckoning and was not agreed even in antiquity. Various ancient scholars each assumed a different date. Each computed his date by adding a different number of years of kingly rule from the foundation of the city to his estimate of the date of the foundation of the republic. This, in turn, was presumably computed by counting back over the yearly lists of magistrates. There may have been traditions about the intervals between certain events in early Roman history, but the frequently accepted reckoning of 244 years of kingly rule seems to be a calculation based only on the conventional 35-year generation for the rule of the seven legendary kings. Polybius claimed that the dating of the first republican consulship to 508/507 BC could be substantiated by an extant copy of a contemporary treaty. Combined with the traditional kingly period, this would give a foundation date of 751–750, reckoned inclusively, and 752–751, exclusively (Cato's date). The chronological scheme worked out by Varro added two years of nonconsular rule, thus the foundation of Rome was put in 754/753 and the beginning of the republic in 510/509. Varro's dates became standard for later Romans and are sometimes also used by modern scholars in a purely conventional sense. But it remains uncertain whether the dating depended on the magistrate lists or whether these were "doctored" to synchronize with given dates or intervals, whether these were traditional or calculated in some other way. Anomalies such as Livy's five-year anarchy 15 years after the Gallic invasion, Diodorus' repetition of magistrates' names, and the "dictator years" in the lists are perhaps attempts to synchronize the various pedigrees.

Contribution of archaeology. Archaeology can provide many dates useful to the detailed study of Roman history, especially from coins and inscriptions, but, for the general scheme of early chronology, its value is largely negative. It shows, for example, that Rome evolved over a lengthy period and was not really "founded," though a "foundation" date might perhaps refer to the first common celebration

of the Septimontium, or festival of the seven hills; again, if that dating is dependent on the seven kings, archaeology shows that the tradition about them, though it may preserve genuine names and events, is largely legendary.

Datings after the 1st century BC. In this better documented period, datings to consul years, or later to the years of tribunician power of the emperors, are normally intelligible, despite a few notorious cruxes, although up to the Julian reform the state of the calendar has always to be taken into account. In parts of the empire, however, different eras were used—*e.g.,* that of the Seleucids—and from the 4th century AD dates were often calculated in terms of the years of the indiction, a 15-year cycle connected with the levying of taxes, a method that continued in use for many centuries in spite of difficulties, such as lack of synchronization among the various provinces.

(Ed.)

The Christian Era

Christian. The Christian Era is the era now in general use throughout the world. Its epoch, or commencement, is January 1, 754 AUC (*ab urbe condita*—"from the foundation of the city [of Rome]"—or *anno urbis conditae*—"in the year of the foundation of the city"). Christ's birth was at first believed to have occurred on the December 25 immediately preceding. Years are reckoned as before or after the Nativity, those before being denoted BC (before Christ) and those after by AD (*anno Domini,* "in the year of the Lord"). Chronologers admit no year zero between 1 BC and AD 1. The precise date of commencing the annual cycle was widely disputed almost until modern times, December 25, January 1, March 25, and Easter day each being favoured in different parts of Europe at different periods.

The Christian Era was invented by Dionysius Exiguus (*c.* AD 500–after 525), a monk of Scythian birth resident in Italy; it was a by-product of the dispute that had long vexed the churches as to the correct method of calculating Easter. Many churches, including those in close contact with Rome, followed 95-year tables evolved by Theophilus, bishop of Alexandria, and by his successor, St. Cyril; but some Western churches followed other systems, notably a 532-year cycle prepared for Pope Hilarius (461–468) by Victorius of Aquitaine. In 525, at the request of Pope St. John I, Dionysius Exiguus prepared a modified Alexandrian computation based on Victorius' cycle. He discarded the Alexandrian era of Diocletian, reckoned from AD 284, on the ground that he "did not wish to perpetuate the name of the Great Persecutor, but rather to number the years from the Incarnation of Our Lord Jesus Christ."

Somehow Dionysius reckoned the birth of Christ to have occurred in 753 AUC; but the Gospels state that Christ was born under Herod the Great—*i.e.,* at the latest in 750 AUC. Dionysius' dating was questioned by the English saint Bede in the 8th century and rejected outright by the German monk Regino of Prüm at the end of the 9th. Nevertheless, it has continued in use to the present day, and, as a result, the Nativity is reckoned to have taken place before the start of the Christian Era.

The new chronology was not regarded as a major discovery by its author; Dionysius' own letters are all dated by the indiction (see below). The use of the Christian Era spread through the employment of his new Easter tables. In England the Christian Era was adopted with the tables at the Synod of Whitby in 664. But it was the use, above all by Bede, of the margins of the tables for preserving annalistic notices and the consequent juxtaposition of historical writing with calendrical computations that popularized the new era. Outside Italy it is first found in England (in a charter of 676) and shortly afterward in Spain and Gaul. It was not quickly adopted in royal diplomas and other solemn documents, however, and in the papal chancery it did not replace the indiction until the time of John XIII (965–972). The Christian Era did not become general in Europe until the 11th century; in most of Spain it was not adopted until the 14th and in the Greek world not until the 15th.

Of the alternative chronologies used by Christians, the most important were: (1) the indiction, (2) the Era of Spain, and (3) the Era of the Passion. The indiction was a cycle of 15 years originally based on the interval between imperial tax assessments but during the Middle Ages always reckoned from the accession of Constantine, in 312. Years were given according to their place in the cycle of 15, the number of the indiction itself being ignored. This chronology was the most widespread in the early Middle Ages, but its use diminished rapidly in the 13th century, although public notaries continued to use it until the 16th. The Era of Spain was based on an Easter cycle that began on January 1, 716 AUC (38 BC), marking the completion of the Roman conquest of Spain. First recorded in the 5th century, it was in general use in Visigothic Spain of the 6th and 7th centuries and, after the Arab invasions, in the unconquered Christian kingdoms in the north of the Iberian Peninsula. It was abolished, in favour of the Era of the Incarnation, in Catalonia in 1180, in Aragon in 1350, in Castile in 1383, and in Portugal in 1422. The Era of the Passion, commencing 33 years after that of the Incarnation, enjoyed a short vogue, mainly in 11th-century France.

(Ed.)

Muslim. Unlike earlier chronological systems in use before Islām, Islāmic chronology was instituted so soon after the event that was to be the beginning of the Muslim era that no serious problems were encountered in its application. According to the most reliable authorities, it was ʿUmar I, the second caliph (reigned 634–644), who introduced the era used by the Muslim world. When his attention was drawn by Abū Mūsā al-Ashʿarī to the fact that his letters were not dated, ʿUmar consulted with men at Medina and then ordered that the year of the *hijrah* (hegira), the Prophet's flight from Mecca to Medina, be taken as the beginning of an era for the Muslim state and community. According to the Muslim calendar, the *hijrah* took place on 8 Rabīʿ I, which corresponds to September 20, 622 (AD), in the Julian calendar. But, as Muḥarram had been already accepted as the first month of the lunar year, ʿUmar ordered that (Friday) 1 Muḥarram (July 16, 622) be the beginning of the reckoning. It is generally accepted that this was done in AH 17 (*anno Hegirae,* "in the year of the Hegira").

The era of the *hijrah*

There are a few points in connection with this that deserve mention: first, there is no real agreement on the exact date of the *hijrah*—other dates given include 2 and 12 Rabīʿ I; second, the year in which ʿUmar issued the order is a point of contention—the years 16 and 17 are sometimes given; third, some people have ascribed the use of the chronology to the Prophet himself. According to some sources, the *hijrah* date was first used by Yaʿlā ibn Umayyah, Abū Bakr's governor in Yemen. This sounds somewhat plausible because Yemenis were probably used to affixing dates to their documents. There is, however, a consensus among workers in the field that 8 Rabīʿ I was the day of the *hijrah,* that ʿUmar instituted the use of the date for the new era, and that this was done in AH 17. The choice of the *hijrah* as the beginning of the epoch has two reasons. On the one hand, its date had been fixed; on the other, ʿUmar and his advisers must have recognized the importance of the migration—Islām had become, as a result, a religion and a state.

Before the introduction of the new epoch, the Arabs had been acquainted with chronologies used by their neighbours, the Seleucids and the Persians. In Yemen the practice of dating had been perfected to the extent that inscriptions show the day, the month, and the year. In Mecca the "year of the Elephant," supposedly coinciding with the birth of the Prophet, had been in use. For the period between the migration and the institution of the new epoch, the Muslims of Medina resorted to naming the year after local events—"the year of the order of fighting" and "the year of the earthquake," etc.

The lunar year was adopted by the Muslims for the new chronology. In this there was hardly any innovation insofar as Arabia was concerned.

The chronology introduced by ʿUmar was adopted throughout the Muslim world, although earlier epochs continued in use in outlying provinces. Muslim historians, annalists, and chroniclers met with difficulties when writing their books on pre-Islāmic history. No practice had as yet developed for pre-*hijrah* dating; therefore, when writing about the history of various lands in pre-Islāmic times,

authors resorted to the use of chronologies previously in existence there (*e.g.,* Persian, Indian, Seleucid, Alexandrian). For the histories of the area under Islām, writers used only Muslim chronology, while non-Muslim authors (*e.g.,* Bar Hebraeus) used the Seleucid and the *hijrah* dates when discussing events pertaining to provinces that had been Byzantine and therefore still had fairly large groups of Christians.

The era of the *hijrah* is in official use in Saudi Arabia, the two Yemens, and in the Persian Gulf area. Egypt, Syria, Jordan, Morocco, Algeria, Libya, and Tunisia use both the Muslim and the Christian eras. Many Muslim countries, such as Turkey, Nigeria, and Pakistan, use the Christian Era.

Variants of the *hijrah* era

Within the general uniformity of applying the *hijrah* era proper, there existed differences, some of which were the result of earlier pre-Islāmic practices; others were the result of continuous contacts of Muslim countries with their European neighbours, with whom they had economic as well as political relations. An example of the former was the work of the ʿAbbāsid caliph al-Muʿtaḍid, who brought the Nowrūz (Persian New Year's Day) back to date in keeping with the agricultural activities of the community. Maḥmūd Ghāzān introduced the Khānian era in Persia in AH 701, which was a reversion to the regnal chronologies of antiquity. It continued in use for some generations, then the ordinary *hijrah* era was reintroduced. A similar step was taken by Akbar when he established the Ilāhī era, which began on Rabīʿ II 963 (February 13, 1556), the date of his accession; the years were solar.

Two Muslim countries, Turkey and Iran, introduced more drastic changes into their chronology because of European influences.

In Turkey the Julian calendar was adopted in AH 1088 (AD 1676–77) and used solar months with *hijrah* dating. The year was officially called the Ottoman fiscal year but was popularly known as the *marti* year, after *mart* (Turkish for March), which was the beginning of the year. Under Mustafa Kemal Atatürk, the Gregorian calendar and the Christian Era were officially adopted in Turkey (1929). Iran also adopted a solar year; the names of the months in its calendar are Persian, and the era is still that of the *hijrah.* (N.A.Z.)

Pre-Columbian American. *Maya and Mexican.* The lowland Maya had a 365-day year formed of 18 "months." Each month consisted of 20 days, plus five "nameless" days, which the Maya considered an extremely dangerous and unlucky period and during which activities were kept to a minimum. Leap days were not intercalated.

Reckoning was not by those years but by *tuns* (360 days) and their multiples of 20: *katuns* (20 *tuns*), *baktuns* (400 *tuns*), *pictuns* (8,000 *tuns*), *calabtuns* (160,000 *tuns*), and *kinchiltuns* (3,200,000). In practice, the last three were seldom used. The *tun* comprised 18 *uinals*, each of 20 *kins* (days), but these did not coincide with the equivalent divisions of the 365-day year. The Maya normally carved or wrote these in descending order; students transcribe them in Arabic numerals—*e.g.,* 9.10.6.5.9 represents nine *baktuns*, ten *katuns*, six *tuns*, five *uinals*, nine *kins*.

With this system, current dates were related to the start of the Maya era, which, because of the Maya system of reentering cycles, marked both the end of 13 *baktuns* (written 13.0.0.0.0) and the start of another cycle of *baktuns* and perhaps commemorated a re-creation of the world, the *baktun* about to enter being numbered 1, not 14. Because of the construction of the calendar, this start of the era happened to be day 4 Ahau falling on the eighth day of the month Cumku.

Such reckonings are called Initial Series, or Long Counts, the former because they usually stand at the start of an inscription (see CALENDAR). For example, the combination day 8 Muluc, falling on second of Zip (third month), recurs every 52 years, but the Initial Series (here 9.10.6.5.9 8 Muluc 2 Zip) pinpoints its position. The next occurrence, 52 years later, would be 9.12.19.0.9 8 Muluc 2 Zip. Each unit had its own glyph (or symbolic character), with appropriate number (normally a dot for 1 and bar for 5) attached.

A shorter dating system was by "Period Endings"—that is, by recording the ending of the current *baktun, katun,* or *tun.* Thus, day 13 Ahau and month position 13 Muan with 13 *tuns* added is an abbreviation of 9.17.13.0.0 13 Ahau 13 Muan, a combination that will not repeat for over 900 years (949 *tuns*). A still shorter but less precise method was to give the day and its number ending the current *katun.*

Several Maya dates were commonly linked to Initial Series or Period Endings by series of additions or subtractions—a glyph signifying count indicated forward or backward by secondary attachments.

Dates were normally reckoned from the 4 Ahau 8 Cumku base, nearly 4,000 years before most inscriptions, but some calculations ranged far into the past and a few into the distant future. One reaches backward nearly 1,250,000 years, but the deepest probings of eternity are embodied in texts that seemingly record positions respectively 90,000,000 and 400,000,000 years ago. Although the interpretation of these last computations is disputable, the Maya certainly thought in millions of years a millennium before Europe discarded the view that the world was only some 6,000 years old.

Maya conception of time

The Maya conceived of time as a journey through eternity in which each deified number—all time periods and their numbers were gods—carried his period on his back supported by a tump line. Each evening the procession rested. Next morning, carriers whose period was completed were replaced. For instance, if the *uinal* and *kin* numbers were 15 and 19 respectively, the new carriers would be the deified 16 and 0 (the latter because *kin* numbers go no higher than 19). Other period numbers would journey on until it came time to change the *tun* carrier. Much ritual and imagery grew out of this concept of the march of time; sculpture illustrates bearers lowering their burdens at journey's end.

Correlation of the Maya calendar with ours depends on several factors. First, the 260-day almanac still functions in some Maya villages in the Guatemalan highlands. As there is excellent evidence it has neither gained nor lost a day since the Spanish conquest, despite strong Spanish efforts to suppress it, one may reasonably assume no break under the more favourable pre-Columbian conditions. Lunar and other data support such a view. Second, month positions in Yucatán and southern Petén at the Spanish conquest also are reliably correlated to the day with the present Western calendar. Third, the combined day and month parts of the Maya calendar are in day-for-day agreement with the present Western calendar within a 52-year span (after that given day and month positions repeat). The *katun* (specifically, 13 Ahau) current at the Spanish conquest is, however, known, thereby fixing any day and month position in a longer range of 260 years because a named *katun* repeats only after 260 *tuns*. Those conditions produce a correlation of the two calendars that is either correct to the day or is 260 or even 520 years wrong, since historical evidence does not specify which particular *katun* 13 Ahau coincided with the Spaniards' arrival. Fourth, such factors as astronomy (Maya records of heliacal risings of Venus and of many dates with moon age stated), pottery sequences, architectural changes (less reliable), and data from neighbouring areas govern choice of the applicable *katun* 13 Ahau. Weight of evidence led to wide acceptance of the Goodman–Martínez–Thompson correlation that equates 13.0.0.0.0 4 Ahau 8 Cumku, start of the Maya era, with August 10, 3114 BC, and the Classic period with AD 300 to 900. Fifth, when the carbon-14 dating technique was first applied to the problem, various difficulties attendant on the use of new techniques and failure to take into account that a tree dies year by year from its centre outward (so that a sample from the core might give a date well over a century before felling) distorted readings, producing results favourable to the correlation making Maya dates 260 years earlier. Now, with better technique and averaging of many "runs" of samples of latest growth from beams at Tikal with secure Maya dates, carbon-14 readings overwhelmingly support the Goodman–Martínez–Thompson correlation.

The only other Middle American calendar with a known era is that of the Cakchiquel of highland Guatemala.

The system was vigesimal: *kih,* day; *uinak,* 20 days; *a,* 400 days; and *may,* 8,000 days. The 400-day "year" ran concurrently with the 260-day almanac, which, in turn, synchronized with all other Maya almanacs. Like the 360-day *tun* of the lowlands, the 400-day *a* was the counting unit, for reckoning was always in multiples of the *a,* never by days, as in our Julian calendar. *May* signifies twenty, and is so named because it comprised 20*a.* At the arrival of the Spaniards, reckoning was from a revolt in AD 1493. Earlier eras may be postulated, but inscribed calendrical texts are lacking in Cakchiquel territory.

Aztec. The Aztec and related peoples of central Mexico employed the cycle of 52 years, constructed, like its Maya equivalent, of concurrent 365-day years and 260-day cycles, any position of the former coinciding with a given position of the latter only at 52-year intervals. Again leap days were not used. At completion of the 52 years, known as "binding of the years," elaborate ceremonies were held to avert destruction of the world expected on that occasion. The last occurrence before the Spanish conquest was in AD 1507. Although the last creation of the world was designated by a day name, neither that nor any other was in general use in central Mexico as the start of an era. Aztec reckoning is normally from their arrival in the Valley of Mexico, supposedly the year 1 Flint (AD 1168).

There is much confusion in placing events in Mexican history because no system of distinguishing one 52-year cycle from another was employed except by writing every year glyph throughout the period covered, a clumsy arrangement. Each year was named for either its last day (omitting the five-day unlucky period) or for the last day of the fifth month (both choices have distinguished supporters). In either case, only four day names (House, Rabbit, Cane, and Flint), each with its accompanying numeral, could designate a year. The Spanish conqueror Hernán Cortés seized the Aztec capital in 1521, year 3 House, but some past event, also assigned to a year 3 House but unlocated in a full sequence of years, might refer to AD 1261, 1313, or 1365, etc. Month positions were rarely given in chronological statements.

Peoples of Oaxaca and the Isthmus of Tehuantepec. Pictorial books of the Mixtec of Oaxaca record events in the lives of ruling families covering seven centuries, but, again, happenings are fixed only by the day on which each occurred and the year in which the day fell. Sequence is usually clear, but at times there is doubt as to which 52-year period is meant when parenthetical material, such as life histories of secondary characters, is inserted.

No era is recognizable. A clouded entry concerning the descent to Earth of the Sun and Venus, perhaps assignable to AD 794, is a logical starting point, but other entries are earlier.

Little is known of the calendar of the Zapotec, neighbours of the Mixtec. Years began on a different set of days, and glyphs differ from those of Mixtec and Aztec. Months are not recorded on monuments, which are numerous, and no chronological system has survived. Most Zapotec texts are early.

Rare inscriptions in western Chiapas, southern Veracruz, and the Guatemalan Pacific coast resemble the abbreviated lowland Maya Initial Series used in script and on a single sculpture in that numerical bars and dots are in a vertical column with period glyphs and month signs suppressed, clearly place numeration, that is, the value of each unit was shown by its position in the column. The linguistic affiliation of their sculptors is unknown.

All texts are either fragmentary or damaged; the two complete ones, unlike Maya Initial Series, open with days signs (and different ones at that). If, as one may reasonably assume, the series of bars and dots departed from those day signs, a fixed era is questionable. Nevertheless, some scholars postulate use of the Maya era (13.0.0.0.0 4 Ahau 8 Cumku). This little understood system may have been ancestral to the Maya Initial Series, the Maya perhaps developing a fixed era, for they alone seem to have been interested in an exact chronological system.

(E.Th.)

DIPLOMATICS

Diplomatics, broadly speaking, is the study of documents. The term is derived from the Greek word *diploma,* meaning "doubled" or "folded." Besides the documents of legal and administrative import with which it is properly concerned, diplomatics also includes the study of other records such as bills, reports, cartularies, registers, and rolls. Diplomatics is therefore a basic and not simply an auxiliary historical science. This article deals with its development and practice in the Roman Empire and in Europe. During Roman antiquity certain documents containing different sorts of authorizations were engraved on a bronze diptych and then folded and sealed, in order to keep the contents secret—hence the term diploma. Rarely found during the Middle Ages, the word was used by the Renaissance Humanists to denote formal documents of ancient rulers. The interest in and description of such documents came to be called *res diplomatica* after the famous 17th-century work *De Re Diplomatica Libri VI,* by Jean Mabillon, a member of the scholarly Benedictine congregation of Saint-Maur. Mabillon's work first made the study of old documents a reputable science. The major task of diplomatics is to distinguish between genuine and false documents, and this involves detailed examination of their external and internal features. Diplomatic studies have been applied mainly to Western documents, usually medieval ones, because it requires less specialist training to analyze more recent documents.

History of the study of documents. *Medieval and Renaissance work.* The forging of documents took place on a vast scale during the earlier Middle Ages, partly because wars and disturbances so frequently upset possession and also because the increasing use of written records made it necessary for those whose title was, in fact, perfectly good in old unwritten "customary" law to give it written substantiation. Thus forgeries, partly intentionally honest, partly dishonest, occurred frequently, despite the fact that the Germanic tribes that settled in western Europe inherited, with other aspects of Roman law, the concept of forgery as a felony, which was soon also reinforced by the church's canon law. This legal concept of forgery was, however, mainly applied to cases concerning property or inheritance; and literary forgeries, such as the famous Donation of Constantine, which purported to be the gift by the Roman emperor Constantine I the Great (died 337) to Pope Sylvester I of spiritual primacy throughout the church and of temporal power in Italy, were not concerned. Serious critical efforts to detect forgery did not begin in the Middle Ages, although obvious forgeries might be challenged in the course of a dispute. As early as the 6th century, the Merovingian king Childebert II declared a charter recording the gift of land from himself to the Bishop of Reims a forgery on the simple ground that the royal official denied the signature on it to be his. Pope Innocent III (1198–1216) tried to establish infallible criteria for the detection of fraudulent papal documents, but knowledge of earlier documentary forms was totally inadequate. In the Renaissance the Humanists began to use philological and technical criteria; on these grounds Lorenzo Valla authoritatively pronounced the Donation of Constantine to be a forgery, though authenticity had already been questioned.

The forging of documents

Post-Renaissance scholarship. Three events in the 17th century forced the development of more sophisticated standards of evaluation. The Thirty Years' War in Germany led to endless legal conflicts, and in France the nobility engaged in a concerted action known as the *bella diplomatica* ("diplomatic wars") to assert their ancient privileges against royal absolutism. The decisive impetus, however, came from a much more particularist dispute. Daniel van Papenbroeck, a member of the Jesuit commission known as the Bollandists (from another member, Jean Bolland), which was charged with the publication of the *Acta Sanctorum* ("Acts of the Saints,"), finding that some monastic documents he inspected were forgeries, assumed (1675) that this was true of almost all early-medieval documents. Since most of the monasteries with which the documents had been concerned were of the Benedictine Order, the Benedictines resented the sugges-

tion, and Mabillon undertook to refute it. In his *De Re Diplomatica* (1681), Mabillon set out the fundamental principles of the science of verifying documents; Papenbroeck soon afterward acknowledged the correctness of his tenets. Nearly a century later, René-Prosper Tassin and Charles-François Toustain published their six-volume *Nouveau traité de diplomatique* (1750–65; "New Treatise on Diplomatic"), a work that surpassed Mabillon's only in its greater wealth of material. Another important event in the history of the science of diplomatics was the founding of the École des Chartes (an institute for the training of French archivists) in Paris in 1821. During the next decades important collections of early-medieval French documents were printed in the *Recueil des actes* by a variety of eminent editors. But the greatest advances were made by German and Austrian scholars, among whom Julius von Ficker investigated the differentiation between *actum* and *datum* (that is, between verbal legal procedure and its formal documentation), and Theodor von Sickel defined a basic technique of studying and comparing the script of charters and thus of identifying the individual notaries or scribes. The diplomas of the Carolingian and the German kings and emperors were edited in the series of the *Monumenta Germaniae Historica,* by members of the Institut für österreichische Geschichtsforschung (Institute of Austrian History Research), established by Sickel in 1854. Meanwhile, the *Regesta,* comprising short, synoptical condensations of the contents of papal documents down to 1198, published by Philipp Jaffé in 1851, gave a decisive momentum to the study of the papal chancery, while August Potthast covered the period from 1198 to 1304. Prominent scholars in the research of papal records in Germany at the beginning of the 20th century were Michael Tangl, Rudolf von Heckel, and, particularly, Paul Fridolin Kehr. In comparison with the amount of work done in France and Germany, historical scholarship in England long paid relatively little attention to legal, as opposed to literary, records. Although John Mitchell Kemble published his collection of Anglo-Saxon documents, the *Codex Diplomaticus Aevi Saxonici* (1839–48), an extensive study of Anglo-Saxon and Norman legal and administrative documents was delayed until the 20th century. Since then notable contributions have been made by scholars such as Helen Cam, H.W.C. Davis, Vivian Hunter Galbraith, Frank M. Stenton, Dorothy Whitelock, David Charles Douglas, and many others. Christopher Robert Cheney has made important contributions to the research of papal documents. In Italy Luigi Schiaparelli made vital contributions to the study of Lombard documents. From the 19th century, some study of documents has formed part of the medieval-history curriculum in most European universities.

ack of udy in ngland

Diplomatic method. *Types of documents.* Documents that have been preserved are either originals, drafts, or copies. Originals, of which many survive, are formal documents drawn up on the order of the sender or donor, and they were designated to serve the recipient or beneficiary as evidence of the transaction recorded. Handwritten copies of documents, made either before or after the deed was actually executed (sealed), are not classified as originals. If made before an "original," they were in fact rough drafts of it; if made afterward, they were copies. The particularly Anglo-Saxon method of chirography, however, gave the possibility of producing several "originals." By this process two or more specimens of a document were written on the same page of the vellum sheet, and the free space between the texts was filled in with the word *chyrographum* ("handwriting") or other words and symbols. Then the sheet was cut irregularly right through these words or symbols; the originals thus separated could later be reassembled, an exact fit being complete proof of authenticity. But to provide documents having the force of "originals," copies of the original were usually made and formally certified as such, by public notaries, or by high ecclesiastical or secular dignitaries. Copies certified in this way were accorded the same legal value as the originals. In practice, lack of critical judgment on the part of the certifiers often led to the certification of forged records. In documents known as transumpts, which recited earlier documents or charters as part of their text, it often happened that the earlier document was forged, but, being included in the new, it received validation. The original documents and copies considered above were issued at the request of the recipient or beneficiary or of his legal heir. It also happened quite often that the sender or donor wished for various reasons to retain a record of the documents issued by him. The chanceries (record offices) of secular rulers or great ecclesiastics therefore kept copies of outgoing documents in registers, and often of incoming documents, too. The popes were among the first to adopt the old Roman practice of keeping registers; although nearly all the earlier ones have been lost, an almost uninterrupted series of papal registers is extant from the pontificate of Innocent III onward. An important group of registers are the rolls kept by the medieval kings of England; the earliest extant rolls date from the 12th century. The keeping of registers in the chanceries of the French kings began about the year 1200, in Aragon about 1215, in Sicily under the Hohenstaufen emperor Frederick II (died 1250), and in the German imperial chancery from the early 14th century. Another manner of studying documents is in the formula books of the various chanceries. Notaries drawing up the various forms of medieval documents did not usually compose each new text afresh but, rather, copied from books in which such text formulas had been collected, a practice that can be traced back to Roman procedure. These model texts frequently contained only the legally relevant passages, while the individually applicable parts, such as names, figures, and dates, were either abridged or totally omitted. During the time of the Frankish kings, important collections were made, such as the *Formulae Marculfi* (early 8th century) and the *Formulae imperiales* (828–832). Significant collections of formulas serving as models for papal documents have been preserved from the 13th century.

Use of formulas

Classification of documents. The documents of the Middle Ages are usually classified under two groups: public documents, which are those of emperors, kings, and popes, and private documents, which comprise all others. Another way of classifying documents is according to whether they are evidentiary or dispositive. The former merely record a valid legal act already executed orally, while the actual issuing of the latter forms in itself the legal act. This distinction, found among Roman documents from the 3rd century AD onward, gradually ceased to exist after the early Middle Ages. After the collapse of the Carolingian empire in the 9th century, private documents lost much of their function and were replaced by simple memorandums about legal acts and the witnesses to them. It was not until the late 11th and early 12th centuries that sealed charters of high secular or ecclesiastical dignitaries were again gradually considered as dispositive. Papal documents can be classified mainly as either letters or privileges, and royal documents can be classified as diplomas or mandates. Privileges and diplomas give evidence of legal transactions designed to be of long duration or even of permanent effect, while mandates and many papal letters contain commands.

Physical appearance of documents. Documents were written on a variety of material. In antiquity there were documents of stone, metal, wax, papyrus, and, occasionally, of parchment, but only papyrus and parchment (and, very occasionally, wax) were used during the Middle Ages. From the 12th to the 13th centuries, paper also was sometimes available. Papyrus, made from the stem of the papyrus plant, was produced mainly in Egypt; after the Arab conquest of Egypt in the 7th century, the import of papyrus to Europe became difficult. The Merovingian kings wrote their documents on papyrus until the second half of the 7th century, and the popes did so until far into the 11th century. North of the Alps papyrus had finally disappeared by the 8th century, when it was replaced by parchment. Parchment was made from animal hides and was thus easier to obtain. In southern Europe it was made mainly from sheep and goat hides; the insides of the skin were thoroughly smoothed and calcined, while the hairy sides were left rougher. In central and northern Europe, parchment was usually made from calves' skins,

Introduction of parchment

and both sides of the hides were thoroughly smoothed and calcined. Paper came originally from China. During the 8th century AD, it spread to the Arab world and from thence to Byzantium, where it was manufactured from linen and was used from the 11th to the 13th centuries for imperial documents. After that time ordinary paper was used in the Byzantine Empire. In the West the use of paper, most common at first in southern Italy and Spain, had begun to spread by the beginning of the 12th century. Germany and southern France began to import paper from Spain and Italy in the 13th century, and soon afterward it had reached England by way of Bordeaux. But paper did not altogether replace parchment, which long remained in use, especially for solemn documents. The medium for writing was ink, generally a mixture of oak gall and copper vitriol. Originally black, ink made north of the Alps sometimes shows a reddish-brown hue, while that made in Italy may contain tinges of brown and yellow. Over the centuries most of these colours have lightened as a result of atmospheric conditions. The Byzantine emperors used purple ink for their signatures. This custom was occasionally taken over by the Lombard rulers of Italy and, later, by the Norman kings of Sicily. Another custom of Byzantine origin is the use of gold lettering.

Throughout the entire Roman Empire, the language used in documents was primarily Latin. Greek was also used, and, during the latter part of the 6th century AD, it slowly superseded Latin in the East. From then onward, Greek was the language of Byzantine documents until the end of the Byzantine Empire (1453). In the West, the collapse of the empire and the establishment of barbarian kingdoms led to a vulgarization of Latin, written as well as spoken.

Latin has always been used for papal documents and for most public and private charters, and it was used for international documents well into post-Renaissance times, until it was superseded by French as the language of diplomacy. In public and private documents, use of the vernacular alongside Latin gradually developed. Apart from its early and unique appearance in the documents of the Anglo-Saxons in England, no vernacular was used in charters before the 12th century. At the Norman Conquest (1066), use of Anglo-Saxon in English documents soon stopped, and no more vernacular was used there until some Norman French was introduced in the 13th century, and Middle English in the 15th century. There was an increasing use of the vernacular in Italian and French documents from the 12th century and in Germany from the 13th; but in medieval times Latin was never outstripped by the vernacular.

Use of vernacular

A correct assessment of the hand in which it was written is vital to ascertaining the provenance and authenticity of a document. Thus, the knowledge of paleography, different styles of ancient writing, is a skill essential to diplomatics. The broad basis of such knowledge begins with acquaintance with the general styles of writing current at particular times and places. This varied with the way the pen was held; whether the writing was cursive or had the letters formed separately; whether it was majuscule, all the letters being contained between a single pair of horizontal lines, or minuscule, with parts of the letters extending above and below the lines. There is a further distinction between what is called book hand and the business, or court, hand at one time used for documents.

In Europe the Roman capital letters, distinguished as rustic or square, uncial, and Roman majuscule and minuscule cursive, influenced all subsequent writing in the West. The Roman curial style (from the Curia, or papal court), used in the papal chancery until the 12th century, was a derivation of late Roman minuscule cursive. After the disintegration of the Western Empire, the Merovingian Franks used a Roman provincial script for their documents. Distinctive forms developed elsewhere, in Visigothic Spain and in Ireland. The Irish script, a half uncial (uncials are rounded letters) and a minuscule script, spread to Anglo-Saxon England and thence to the European continent. Under the Carolingian rulers, a particularly clear and attractive minuscule book hand (Caroline minuscule) was developed; modifications of this gradually became used in documents and eventually spread also to Italy, England, and Spain. A "Gothic," more pointed form of script developed since the 11th century in northern France and soon spread all over Europe, so that writing became more spidery in appearance. In the early years of the Renaissance, Italian scholars such as Poggio (Poggio Bracciolini) and Niccolò Niccoli developed a minuscule based on the Carolingian, and variants of this style were used by the Venetian Aldus Manutius and other pioneers of printing.

Abbreviations were used in both documents and books. Again, their particular characteristics would contribute to a correct assessment of the probable date and provenance of a document. Roughly two types were used: the suspension, involving the writing of only the first letter or syllable of a word; and the contraction, used first for Hebrew and Christian sacred names, the writing of only the first and last letter or letters of a word or syllable. The Carolingians sometimes used Tironian notes, a form of shorthand devised by Tiro, a freedman of the Roman orator Cicero.

From Roman times the two most important methods of validating documents were by appending the signature or the seal of the sender or promulgator. The practice of using seals for this purpose (and not merely to close a document) was carried over from imperial usage and, by the 8th century, was current among the Lombards and other Germanic tribes in western Europe. Until about the 8th century, the signature of the Merovingian ruler or his delegate was also required for the validation of public documents, but thereafter the seal alone, together with the recognition by a high chancery official, was held sufficient, the king's signature dwindling into a monogram or mere stroke (the "stroke of execution"). This change was probably accelerated because many medieval kings could not write. Thus, in England, King John sealed, and did not sign, the Magna Carta.

Seals were made of wax or of metal; if the latter, they were called bulls (hence, the use of this term for a certain group of papal documents). The Byzantine emperors used gold seals for their documents; Byzantine officials and ecclesiastics used lead and silver for their bulls. Papal seals were of lead or gold. Wax seals were increasingly used from about the 11th and 12th centuries, and wax was also used for the impression when, later, less formal documents were validated by use of the signet or privy seal. Seals could be two-sided, suspended from the document, or impressed upon it.

Seals

Form and content of documents. Normally, each document was divided into three distinct parts: the introduction (protocol), the main text (context), and the concluding formulas or final protocol. There were various subdivisions, and not all the parts here mentioned are necessarily found in every document. The introduction comprises, first, the invocation (*invocatio*) of God, either by name or through a symbol such as the cross; second, the superscription (*intitulatio*), giving the name and title of the sender; and third, the address (*inscriptio*), naming those to whom the document is directed, usually followed by a formula of greeting (*salutatio*). The actual text of the document can be divided into a number of parts. The first, known as the *arenga,* expresses in general terms the motive for the issue of the document. The notification (*promulgatio*), briefly explaining the legal purpose of the document, is followed by the *narratio,* or exposition of the particular circumstances involved. In the *dispositio* the donor or promulgator firmly declares his purpose ("I hereby decree" or "I hereby give"); this clause is the vital core of the document, its legal decree of enactment. There usually followed the *sanctio,* a threat of punishment should the enactment be violated. The main text concluded with the *corroboratio,* a statement of the means to be used for validation of the document. The final protocol consisted of subscriptions or lists of names of all those, such as the scribe, who took part in the issue of the document and of witnesses to the enactment. The date and place of issue are given, and the final sentence, the *apprecatio,* is a short prayer for the realization of the contents of the charter. At the bottom of the document, the signs of validation (the recognition, monogram, seal) were then added.

The date given on a document might be either that of legal enactment (*actum*) or that of the issue of the doc-

Problems of chronology

ument recording the (already performed) legal enactment (*datum*). The form in which dates are given in a document is of particular import in determining its provenance and authenticity. A wide variety of practices were followed at different places and times. For instance, days of the month could be given according to the old Roman system of calends, ides, and nones; by continuous counting throughout the month; or by reference to a saint's day. Years might be computed from the presumed time of the creation of the world; by the Roman indiction, a 15-year cycle; by the names of officiating Roman consuls; by regnal years of emperor, king, or pope; or from the birth of Christ. Moreover, there were also a variety of ways to determine when the year began.

Development and characteristics of chanceries. *The Roman and Byzantine empire.* Rulers, all of whom needed to issue directives and edicts, developed writing offices, or chanceries, in which formal documents were drawn up. The Roman imperial chancery, called the Office of Letters (*ab epistulis*), was subdivided into a Greek and a Latin department. In the 5th century four letter offices existed, all under the ultimate control of the *magister officiorum* ("master of offices"): the *scrinium epistolarum* ("letter office") handled mainly foreign, legal, and administrative affairs; the *scrinium libellorum* ("petitions office") handled petitions and investigations; the *scrinium memoriae* ("memorandum office") composed shorter imperial decrees; and the *scrinium dispositionum* dealt with administration. From the 4th century, a group (*schola*) of notaries had come into being, some of whom served the emperor as personal secretaries. Two centuries later a special confidential (*a secretis*) secretary existed. In the Byzantine Empire in the 8th to 9th centuries, the *scrinium epistolarum* and the *scrinium libellorum* merged to form a new department under the *koiaistor* (a high palace official), while the secretaries had all come under the office of the *protoasekretis* (head of the secretaries). An official called the *mystikos* handled the emperor's secret correspondence. In preparing edicts or other laws, the *koiaistor,* after consulting the emperor, made a first draft of the bill, had the official copy drawn up by notaries, and then verified its accuracy before it was validated. From the 9th century onward other high court officials participated in the validation of Byzantine charters.

The important governmental documents of the late Roman and early Byzantine empires include laws, edicts, decrees (imperial decisions concerning civil and penal law), and rescripts (the emperor's replies to inquiries from corporate and administrative bodies or private persons). In the Byzantine era documents concerning more day-to-day affairs can be grouped under the headings of foreign letters, privileges, and administration. Foreign letters include correspondence with other rulers, treaties (regarded not as an agreement between equals but as an act of grace or privilege granted by the emperor, and made out as such), and letters accrediting imperial ambassadors. The most solemn and splendid form of privilege was the *chrysobullos logos,* so named because the word *logos,* meaning the emperor's solemn word, appeared in it three times, picked out in red ink. Written in the carefully embellished chancery script reserved for the emperor's personal documents, the text consists of the usual parts—that is, the *invocatio, intitulatio, inscriptio, arenga, narratio, dispositio, sanctio,* date, and the *subscriptio.* It was sealed with the golden bull.

Imperial privileges

From about the 12th to the mid-14th century, a simplified form, the *chrysobullon sigillion,* was used for privileges of lesser importance. It was not signed by the emperor himself but was held to be validated by the insertion, by the emperor, in red ink of the *menologema,* a statement of month and indiction. It, too, was sealed with a golden bull. The administrative documents of the Byzantine imperial chancery include the *prostagma,* or *horismos,* a plain and short document known since the beginning of the 13th century. If directed to a single person, the document starts out with a short address, but, in all other cases, it begins immediately with the *narratio,* followed by the *dispositio.* The emperor replaced his signature with the *menologema.* Unlike the privileges, this document was not rolled up but, instead, was folded, and then closed by means of a wax seal stamped with the imprint of the imperial signet ring.

In addition to those emanating from the imperial offices, there were other types of documents issued in the Byzantine Empire. These include those issued by despots and imperial officials and, in the ecclesiastical sphere, by patriarchs and bishops. There were also private documents. The documents issued by the despots carried a silver seal, showing their intermediate status between that of imperial documents sealed with the gold seal and that of documents drawn up by imperial officials and sealed with lead bulls. Documents issued by imperial officials were simpler. They lacked the protocol, and the personal signature of the issuing official was written in black ink. The detailed date comprises the *menologema.* The documents of the Byzantine patriarchs are in many respects analogous to the imperial documents and symbolize the high status of the patriarch of Constantinople. They were, however, sealed with lead bulls. Byzantine private documents are almost exclusively notarial instruments. They are immediately recognizable by the crosses marked at the top of the documents. Used in lieu of the signature, the cross was the mark of the sender and contained his name and official function in one of its angles. The document was either signed by witnesses, or at least the cross preceding their names is autograph. Following that is the signature of the notary. These documents were not usually sealed.

The papal chancery. Knowledge about early papal documents is scant because no originals survive from before the 9th century, and extant copies of earlier documents are often much abridged. But it is clear that the popes at first imitated the form of the letters of the Roman emperors. The papal protocol consisted only of the superscription and address and the final protocol of the pope's personal "signature"—not a mention of his name but merely a blessing. Toward the end of the 8th century, it became customary in certain documents to mention in the final clauses the name of the scribe responsible for the drawing up of the document; this was given with the date of issue, indicated by month and indiction, immediately following the subject matter of the document. There followed another clause, the great dating formula, *datum per manus* ("given by the hand of . . ."), naming a high chancery official and giving the date by reference to the regnal years of both emperor and pope. Both were used in documents containing decrees of permanent legal force, which came to be called privileges. Under Pope Leo IX (1049–54), the benediction written by the pope was changed into a monogram not written by him, but his signature was now introduced, placed in a round symbol, the *rota.* By the early 13th century, papal documents had evolved into two distinctive groups: solemn privileges and letters. Solemn privileges can be distinguished by their enlarged letters (*elongata*) of the first line, by the phrase *in perpetuum* ("in perpetuity") at the end of the address, by a threefold *amen* at the end of the text, by use of the *rota,* the pope's signature, the monogram, signatures of the cardinals, and by the *datum per manus.* Among letters, those whose bull was fastened on silken cords (*litterae cum serico*) brought some benefit to the recipient, while those with bulls fastened on a hempen cord (*litterae cum filo canapis*) contained either orders or the papal delegation in a dispute.

Great dating formula

The number of solemn privileges began to decline from the mid-13th century, and eventually they were completely discontinued, their function being partly taken over by the *litterae cum serico,* which became increasingly elaborate in form. A new type of document also developed, the papal bull, distinguishable primarily by its use of formulas such as *ad perpetuam rei memoriam* ("that the matter may be perpetually known") in the superscription. Yet another new papal document appeared at the end of the 14th century, the brief (*breve*), used for the popes' private or even secret correspondence. Written not in the chancery but, instead, by papal secretaries (an office dating from about 1338), the briefs were sealed on wax with the imprint of the papal signet ring.

The papal chancery of the 4th to the 8th centuries was similar to the late-Roman imperial chancery. Its notaries (*notarii, scriniarii*), organized in a guild (*schola*), were headed by the *primicerius notariorum* and the *se-*

Chancery officials

cundicerius notariorum (first and second of notaries) and included the especially important notaries of Rome's seven ecclesiastical regions. But, during the 9th century, the *bibliothecarius,* the papal librarian, became the most important chancery official; a little later, various important bishops and dignitaries seem to have acted occasionally as *datarius* (the official named in the *datum per manus* formula). During the mid-11th century, a phase of German influence led to the temporary employment of notaries from the court of the emperor Henry III, who drew up papal privileges according to imperial formulas. A more important and permanent outcome of German influence was the gradual replacement of the *bibliothecarius* by a chancellor as the highest chancery official. The chancellor was invariably a cardinal, and in his absence another cardinal acted in his place as vice chancellor. Lesser chancery personnel still included the seven regional notaries; increasing business involved the use of lesser paid scribes in addition to the established notaries. From the late 11th century, a papal chapel, modelled on those of contemporary emperors and kings, developed, and its staff was often employed in chancery tasks.

From the early 13th century, the vice chancellor became the permanent head of the chancery, the office of chancellor remaining vacant. During that century the vice chancellors were ordinary clerics, who renounced the office if elevated to the cardinalate; thus, the chancery became directly subordinate to the pope himself. Both the numbers and the official standing of the notaries in the chancery, which then functioned entirely separately from the chapel, gradually increased. Higher chancery officials were often distinguished canonists (legal experts), such as Sinibaldo Fieschi (later Pope Innocent IV), Godfrey of Trani, and Richard of Siena. From the beginning of the 14th century, bishops or cardinals filled the office of vice chancellor. During the Great Schism (1378–1417) there were two papal chanceries and two vice chancellors, one in Rome and one in Avignon.

Under Innocent III the procedure of the papal chancery had changed. Letters concerning matters of import to the papal Curia (*de Curia*) were drafted by the pope himself or else by a cardinal, the vice chancellor, or a notary. But the majority of the papal documents were elicited by their recipients, who had first to present to a notary the substance of their petition in a form the text of which largely anticipated the wording of the desired document. Professional proctors attached to the Curia assisted in the drafting and were also responsible for the documents during later stages of the procedure. Once a petition was approved, the notaries or the *abbreviatores* drafted a suitable document, drawing on a selection of formula books. After a final copy (engrossment) had been made and checked, it was read, if necessary, to the pope or in a special department of the chancery, the *Audientia litterarum contradictarum.* It was then passed to the Cistercian lay brothers who had charge of the papal bull, sealed, and given to the petitioner, who had had to pay a fee at almost every stage of the proceedings.

Insufficient research has so far been done on the papal chancery during the 14th and 15th centuries. Whereas formerly, when the vice chancellor was absent, one of the notaries had deputized for him, a new official, the *regens cancellariam,* was now created to fulfill this function. The number of notaries increased steadily, and, from the 13th century onward, an increasing number of public notaries worked in the papal administration. In order to distinguish between them and the papal notaries proper, the latter became unofficially known as protonotaries. The notaries were now in charge of the letters of justice, while the letters of grace were handled by the *abbreviatores.* The scribes remained in charge of the engrossments. A *computator,* aided by several assistants, was responsible for collecting fees.

The royal chanceries of medieval western Europe. Of the nations that held power in western Europe after the collapse of the Roman Empire there, the Ostrogoths, who occupied Italy from the late 5th to the mid-6th century, took over the ancient Roman imperial-chancery system in its entirety. Very little is known about the royal documents of the Lombards, their successors in Northern Italy, since not one of them has been preserved in its original form. But Lombard officials in charge of drawing up the documents were still trained in the Roman tradition. As well as *referendarii,* there were notaries who also acted as scribes. It is very likely that all of them were laymen.

Until the 12th century two main types of documents, diplomas and mandates, were produced north of the Alps, in the Merovingian, Carolingian, German, and French royal chanceries. Very little is known about the Merovingian royal chancery and its organization. The names of the scribes are never mentioned in the documents, but they were signed by high chancery officials, the *referendarii.*

When the Merovingian dynasty was supplanted by the Carolingians, chancery procedure changed drastically. In contrast to the Merovingian kings, the first Carolingian king, Pepin the Short, was unable either to read or write. He therefore entrusted the responsibility for the correctness of the royal documents to an official of the court. At about the same time, the task of drawing up documents was taken over by those clerics whose original duty had been to look after the most important relic of the royal court, the coat (*cappa*) of St. Martin of Tours. Collectively named the *capella* (chapel), these clerks were individually called *capellani,* chaplains. This close connection between the court chapel and the chancery existed under the later Carolingians and at the German and French and other royal courts, including that of England. Until well into the 12th century, European chanceries were not bureaucratic offices in the modern sense but, rather, in most cases an assemblage of chaplains suited for the task of issuing documents and usually working under a cleric who was not the head of the chapel. Not all chaplains wrote documents, however, and the chapel and chancery thus remained separate institutions. From the reign of the emperor Louis I the Pious (814–840), the heads of the chancery were not personally involved in writing the documents, a task performed by unnamed and unknown scribes. At first the scribes were indiscriminately designated as either *notarii* or *cancellarii* (higher, Roman provincial officials of the 5th and 6th centuries, who stood at the barriers, *cancelli,* of the council rooms), but, by the 9th century, the title of *cancellarius* was gaining ground and was increasingly applied to the head of the chancery. The 9th century was a period of transition, during which, for a while, the archchaplain, the head of the chapel, became also the head of the clerks who wrote the charters.

The *capella* of St. Martin

Under the Ottonian dynasty, which came to power in the eastern division of the original Carolingian empire early in the 10th century, the German royal chancery developed the organization that was to characterize it throughout the remainder of the Middle Ages. The heads of the chancery were the archchancellors, but the office was entirely honorary and soon came to be automatically held, as far as Germany was concerned, by whoever was archbishop of Mainz. When the German kings or emperors established administrations in Italy, Italian bishops were at first made archchancellors for Italy, but in 1031 the office was attached to the archbishopric of Cologne. From the 11th century, Burgundian bishops were archchancellors for Burgundy, but, in the second half of the 13th century, the archbishop of Trier took over the office.

Imperial chancellors

The actual heads of the chancery were the chancellors. At first there was a chancellor, as well as an archchancellor, for each separate part of the empire—Germany, Italy, and Burgundy—but from 1118 there was only one chancellor for all three kingdoms. But even the chancellors, all of whom were clerics, were rarely involved in the actual composition and engrossing of documents, being usually engaged, as important advisers to the king or emperor, in much weightier matters. They do seem to have been especially concerned, however, with decisions about the granting of charters, and they supervised the work of the scribes or notaries. From among the ranks of these notaries, a group of protonotaries gradually developed after the mid-12th century, as a result of influence from the chancery of the Norman rulers of Sicily. Often called upon to deputize for the chancellor, the protonotaries, from the

late 13th century onward, frequently titled themselves vice chancellors.

From the 12th century onward, the documents issued by the German royal chancery were divided into various classifications. The diploma, by then usually called a privilege, existed in two categories, the solemn and the simple privilege. A solemn privilege included the *invocatio,* the signum and recognition line, and a detailed dating or at least one of these three elements, which were entirely lacking in simple privileges. Gradually, simple privileges merged into documents called mandates; it is not always easy to distinguish between them, but, in general, privileges were concerned with rights in perpetuity, while the mandates dealt mainly with matters of only temporary importance. From the early 14th century, mandates were superseded by the use of letters patent and letters close (open or closed letters). Privileges continued to be sealed with a hanging seal; the seal on letters patent was impressed on the document and was used to seal up letters close.

As the power of the German kings declined during the later Middle Ages, so that of the archchancellors increased, and in the 14th century they attempted to win control of the chancery. But, despite fluctuations in the power struggle, the king retained control of the chancellor, who, by the end of the 15th century, held the title of imperial vice chancellor.

Under the Carolingians and the first Capetians in France, various bishops and archbishops, especially the archbishops of Reims, held the office of royal chancellor. But at that time the office was merely titular, and, by the end of the 11th century, it disappeared entirely. From the 12th century onward, the title of chancellor became reserved to the head of the chancery. These new chancellors became so powerful that in 1185 King Philip II Augustus left the office vacant, and, during almost the whole of the 13th century, the chancery was administered by subordinate officials. Chancellors, often laymen, were appointed again in the 14th century, however, and the office remained important until 1789. As in other parts of Europe, the French chancellor merely directed the work of the notaries, and it was they who were responsible for drawing up the documents. From 1350 onward, the notaries were called secretaries, and both their numbers and their importance steadily increased. From the 15th century, the tremendous expansion of business occupying the Grande Chancellerie led to the establishment of several subsidiary *petites chancelleries,* all issuing royal documents sealed with the king's signet. Until the reign of Henry I (1031–60), the old Frankish type of diploma was issued almost exclusively. Then, gradually, charters in the simpler form of letters began to replace the diplomas, and, during the 13th century, the *lettres patentes* became the common type of document. These lacked the *invocatio,* the monogram, and the signature of the high dignitaries, and they gave the simple form of dating. From the 14th century, two forms of *lettres patentes* existed: the *charte,* which was sealed with a green wax seal hanging on red and green silk cords; and the *lettre patente,* used mainly for administrative mandates, which was sealed with a yellow wax seal on double and single cord. Besides, *lettres closes* were used from the 13th century onward.

The English royal documents of the Anglo-Saxon period (before 1066) can be divided into two large groups: the charters, mostly written in Latin; and the writs, written in Old English. The charters, for the most part concerning grants of land, began with either a verbal or a symbolic *invocatio* (a cross or the monogram XP for Christ). There was an *arenga* but no *intitulatio.* Charters were not sealed, the validation comprising the nonautographed signatures of the king and of ecclesiastic and secular dignitaries. Many of the extant charters of this era are forged or interpolated; the series of those apparently genuine starts shortly after the arrival in Canterbury (669) of the archbishop Theodore, and these show similarities with late-Roman private documents. It therefore seems probable that the Anglo-Saxon charters derive from Italian models brought to England by the Roman missionaries. Most of the charters were apparently written by the recipients. From the 11th century, the charters were gradually superseded by sealed writs, which became the most important type of document in medieval England. At first written in Anglo-Saxon, they were produced in Latin soon after the Norman Conquest. No continental document was anything like them. Written on a narrow strip of parchment, their entire text occupied only a few lines. A symbolic *invocatio* in the form of a cross was followed by the king's name and the address, which contained a salutation clause. There was no *arenga,* and the address was followed by a short description of the bequest, if the writ concerned a grant, or a directive, if the writ was a mandate. The Anglo-Saxon writs had no signatures of witnesses and no date, but after 1066 these elements were added. The dating consisted at first only of a mention of the place of issue, days and month being usually given only toward the end of the 12th century. From the very beginning the writs were supplied with a pendant seal as a means of validation. From the reign of Henry II (1154–89), it is possible to distinguish mandate-like writs of the old type and charter writs, which mainly concerned questions of feudal enfeoffments and confirmations of privileges and were usually more carefully executed than the mandate writs. By the early 13th century, the charter writs had developed into a new form of charter. It contained the *intitulatio,* including all the titles of the king; the general address; the dispositive text, which in most instances was introduced with the expression *sciatis* ("know that"); and the final clauses, consisting of the names of several witnesses and of the *datum per manus* clause mentioning the chancellor and giving the date and place of issue. During the final stage of its development, the great seal of green wax, pendant on silk cords, served as the means of validation. The ordinary writs further evolved into the common-law writs (containing orders to persons mentioned by name), the writs of summons, and, particularly, the letters patent (the latter eventually assuming the function of the charter writs, which disappeared at the end of the 13th century) and the letters close. The letters patent were furnished with a general address. Often introduced with the *sciatis,* the text concerns either limited grants or commissions to royal officials, introduced by the word *precipio* ("I order") or a similar expression.

English writs

The dating begins with a series of witnesses and contains the place of issue, day, month, and year of the king's reign. Occasionally the dating is followed by the words "per N.," which are assumed to designate the household official transmitting to the scribe the royal order to issue the writ. The means of validation was the great seal in white wax, pendant on a single strip of parchment (*simple-queue*) or on a double strip of parchment or silk (*double-queue*). The letters close obtained their name from the fact that they were closed by means of the great seal. They contained either commands or information directed to a single individual or to several persons. Characteristic of these letters are the words *teste me ipso* ("witnessed myself") introducing the regular dating clause. These types of royal documents remained essentially unchanged throughout the Middle Ages and were imitated by both secular and ecclesiastic English magnates and dignitaries.

The English royal chancery grew out of the royal household. As on the Continent, it was at first merely a group of royal chaplains, and, until the Norman Conquest, there was no chancellor at their head. The chancellor was the keeper of the seal but usually took no part in the issuing of documents. During the 12th century the number of scribes was still low, between two and eight. At first this "chancery" travelled about with the king, only in the course of the 13th century establishing a permanent location at Westminster. At first English royal documents were not dispositive but merely evidentiary, confirming a previously arranged, orally discharged legal act, and the king's order for issuing the document was given orally. But, from the late 13th century, a royal secretariat came into existence, which was in charge of relaying to the chancellor, by warrant sealed with the privy seal, the royal order for issuing a writ under the great seal. In many ways the secretariat thus became a competitor of the chancery. In addition, during the first half of the 14th century, the Signet Office was established, so called after the small seal (signet). The king's secretary was also the head of this office. All these

The Signet Office

shifts made the issuing of royal documents increasingly complicated. From the end of the 14th century, the common procedure involved, first, the petitioner submitting a petition to the king. If the king approved of it, his secretary forwarded a warrant carrying the signet to the keeper of the privy seal with the request to send, in his turn, a warrant carrying the privy seal to the chancellor. The chancellor then ordered the issue of the document, which would bear the great seal. So far as the issuing of royal documents is concerned, the fact that the secretariat developed into the secretariat of state is of great significance. The king's secretary (later on, there were two of them) became the centre of the royal administration. It was in his office that the state papers originated, which already under Henry VIII (died 1547) far surpassed in importance the old chancery records. Among the state papers there are in-letters, out-letters, drafts, reports, and schedules. The decline of the rolls (document registers) during the 16th century gave rise to yet another new office, the State Paper Office, headed since 1578 by the clerk of the papers. The second holder of this office, Sir Thomas Wilson, established the division of the state papers into foreign and domestic. As departments of state proliferated during the 18th and 19th centuries, they developed their own archives. In 1838 all the public archives became subject to an official called the master of the rolls.

Smaller European nations usually modelled their documents on those of the papacy, the empire, France, or England. The influence of papal letters and privileges can be observed particularly in Aragon, Castile, and Portugal, while German royal diplomas served as models in Bohemia (which was part of the empire), Hungary, and Poland. Because of the close political ties between the two kingdoms, Anglo-Saxon influence can be traced in the seal of the royal Danish documents during the 11th century, but, in the course of the 12th century, royal and princely German documents became the models for Danish as well as Swedish royal documents. Norwegian royal documents were modelled on Anglo-Saxon writs, probably as a result of the influence of English missionaries working in Norway from the early 11th century. The Norwegian writs were drawn up in the vernacular.

The chancery of Norman rulers

The chancery of the Norman rulers of southern Italy and Sicily was highly developed. Influenced by the form of papal documents, the Norman documents comprise mainly privileges, either formal (with *rota,* witnesses, and gold bull) or simple (with *rota,* leaden bull, or wax seal), and mandates. They all have a detailed dateline that includes the name of the chancellor or other high court officials, the number of years since the birth of Christ, the regnal year of the king, and the *apprecatio.* The mandates are more simply executed. They lack the *invocatio* and start out with a simple *intitulatio* and *inscriptio* that ends with a salutation clause. There is neither *arenga* nor *corroboratio,* but there is a command clause in the text. The dating consists only of the place of issue, the day of the month, and the *indictio.* There is no *rota* or signature. The seal is of red wax. The head of the royal chancery of the Norman kingdom of Sicily was the chancellor, a layman who was an influential court official. The notaries who drafted and wrote the documents were also laymen. Because the German Hohenstaufen emperors also ruled in Sicily from 1194 to 1250, Norman chancery practice influenced subsequent German documents. (Pe.He.)

EPIGRAPHY

Epigraphy (from the classical Greek *epigraphein,* "to write upon, incise," and *epigraphē,* "inscription") denotes a branch of scholarship devoted to the study of written matter recorded on hard or durable material. Because such media were exclusive or predominant in many of mankind's earliest civilizations, epigraphy is a prime tool in recovering much of the firsthand record of antiquity. It is thus an essential adjunct of the study of ancient man; it secures and delivers the primary data on which historical and philological disciplines alike depend for their understanding of the recorded past. In a narrower sense, epigraphy is the study of such documents as remains of the written self-expression of early cultures and as communication media in their own right, attesting to the development of visible sign systems and the art of writing as such. Finally, in later periods including our own, where perishable writing media predominate, epigraphy affords insights into the styles and purposes of monumental or otherwise exceptional techniques of written recording.

Uses of epigraphy

Materials and techniques. The delimitation of epigraphy vis-à-vis contiguous and related areas of antiquarian scholarship meets with some ambiguity. In a wide sense, epigraphy concerns itself with the total firsthand transmission of the written remains of ancient civilizations (as opposed to *post-factum* copying) and the nature of the material (*e.g.,* stone, marble, metal, clay, terra-cotta, pottery, wood, wax tablets, papyrus, parchment) and the technique of recording (cutting, carving, engraving, casting, embossing, scratching, painting, drawing, etc.) have mere secondary relevance. Under this maximum definition certain subdisciplines may be included under the overall canopy of epigraphy: notably numismatics, which concerns itself with legends on coins and medals, and papyrology, the study of a special type of perishable record that is normally preserved only in the dry climate of Egypt and in adjacent desert regions. In the case of Egypt, papyrology tends to impinge upon wood and clay media as well, thus leaving mainly stone and metal objects as the concern of epigraphy proper.

In general, however, unless so subdivided, epigraphy encompasses inscriptions at large, be they on primary writing surfaces or on such assorted objects as vases, potsherds, gems, seals, stamps, weights, rings, lamps, and mirrors. A further related discipline is paleography, which concerns itself with the study of scribal hands and styles of writing and has significance for the dating of epigraphic as well as other written documents.

Monumental, archival, and incidental inscriptions

The nature of the materials and techniques used for inscriptions is closely tied to the external purpose of the record itself. Thus, inscriptions may be divided into monumental, archival, and incidental. Monumental inscriptions were intended for enduring display and were therefore, as a rule, executed in lasting material, such as stone or metal. Maximal exposure to mortal eyes need not have been the prime purpose of their originators—*e.g.,* the tomb chambers of Egyptian pharaohs, intended to be sealed forever, had their inner surfaces covered with monumental hieroglyphs; the great Bīsitūn inscription of King Darius I of Persia is on a high rock surface and legible only after precarious rock-climbing or from airborne conveyances. Under this classification may be included also micromonumental inscriptions found on such objects as coins, seals, and rings, meant to endure in their own right.

Archival inscriptions were essentially a feature of those early societies that kept records and that used such materials as have been preserved thanks to their intrinsic, accidental, or incidental durability. Many ancient Near Eastern cultures employed clay tablets for writing, which they fired to insure their soundness. Minoan and Mycenaean archivists in ancient Crete and Greece used perishable temporary clay records that were preserved by unintentional baking in the conflagrations that destroyed their storerooms. Papyrus records from Egypt have survived as a result of climatological chance—mainly low humidity. The official purposes of public display and of archival preservation were sometimes complementary, and therefore coincidental or overlapping matter has been preserved. In some cultures the techniques employed in monumental and archival writing tended to differ (notably in Egypt, where increasingly cursive hieratic or demotic script contrasted sharply with monumental hieroglyphic), and occasionally the language itself would be different (for example, in the Hittite Empire, where the clay tablets in cuneiform employed mainly straight Hittite or Akkadian, whereas the monumental "hieroglyphic" rock inscriptions and seals used a distinct language).

Incidental inscriptions may be defined as those not seriously meant for preservation. They include, for example, wall scrawlings of the graffiti type and casual records that were kept on cheap writing matter such as potsherds (*ostraca*) and scraps of papyrus. Many a city dump of ancient Egypt has yielded a rich harvest for the study of daily life.

Inscriptions as historical source material. In study-

ing the political, administrative, legislative, and dynastic records of extinct civilizations, the modern historian must bring to bear all the evidence at his disposal; and such evidence may vary sharply from one locality and period to another. Historiography in the modern sense—the analytical ordering and interpretation of past institutions and events—is an invention of ancient Greece, and even there it only gradually eschewed the fabulous. In many early societies (*e.g.,* Mesopotamia, Egypt), chronographic records were either annalistic or legendary in kind and sometimes apologistic or propagandistic in purport; and in others (*e.g.,* India) a cyclical world view prevailed, and a lack of feeling for linear time depth precluded an ordered appreciation of the past, leaving only legendry—in such cases only synchronisms with time-anchored events elsewhere allow a precise dating.

Thus, the amount of predigested ancient information bearing on antecedent events may vary from sophisticated literary to scrupulous epigraphic or may be wholly lacking or largely valueless. In the latter instances the historian is almost exclusively dependent for native information on primary documents, and such documents are in most cases inscriptional.

Ancient Mesopotamia. Surviving epigraphic matter from the 3rd and early 2nd millennia BC includes both historical and quasi-historical material. The Sumerian king list is a compilation of names, places, and wholly fabulous dates and exploits, apparently edited to show and promote time-hallowed oneness of kingship in the face of the splintered city-states of the period. The Sargon Chronicle is a piece of literary legendry concentrating on spectacular figures and feats of the past, whereas contemporary royal inscriptions, notably by Sargon I of Akkad and Gudea of Lagash, are historical documents in the proper sense.

Both kinds of texts are preserved also from the Babylonian and Assyrian periods, from the reign of Hammurabi (1792–1750 BC) to the 6th century BC. There are lists of date formulas and year names from Hammurabi's reign and from that of his son Samsuiluna; lists of Assyrian eponymous year names, based on those of dignitaries; the Babylonian king lists, running from Hammurabi through the Kassite era and the Assyrian domination of Babylon to the last flicker of Babylonian self-assertion in the early 6th century BC; the Assyrian king list from Khorsabad, which made good use of earlier compilations; and notably the so-called Synchronistic Chronicle, which juxtaposed the kings of Assyria and Babylonia in the same millennial sequence. Historical documents comprise, above all, the stately sequence of annals by the kings of Assyria, recorded on stone slabs, stelae, foundation markers of buildings, bronze gates, statues, and obelisks and in clay archives (prisms, cylinders, tablets). Starting in the Old Assyrian period, they were especially extensive in the reigns of Tiglath-pileser I (1115–1077 BC), Ashurnasirpal II (883–859 BC), Shalmaneser III (858–824 BC), Adadnirari III (810–783 BC), Tiglath-pileser III (744–727 BC), Shalmaneser V (726–722 BC), Sargon II (721–705 BC), Sennacherib (704–681 BC), Esarhaddon (*c.* 680–669 BC), and Ashurbanipal (668–627 BC).

For all their swaggering bombast and flaunting of deliberate cruelty, the annals provide prime historical source material. The detail of the Assyrian conquest of Syria, Palestine, parts of Asia Minor, Cyprus, Arabia, and Egypt would be spotty indeed without recourse to these annals, for they show the centre of political power, unlike such provincial records as those from contemporary Egypt or the Old Testament.

Legal compilations and law codes also have pride of place in the epigraphic record of ancient Mesopotamia. These form a unique succession, starting in the 3rd millennium BC with that of King Ur-Nammu of the Sumerian 3rd dynasty of Ur (*c.* 2100 BC), continuing with those of the Sumero-Akkadian king Lipit-Ishtar (in Sumerian) and King Bilalama of Eshnunna (in Akkadian) during the interval of the 3rd dynasty of Ur, and the rise of the Amorite dynasty of Hammurabi (*c.* 2000 BC), culminating in the great diorite stela of Hammurabi (*c.* 1750 BC), showing retardation and recrudescence in the Middle Assyrian laws that are found on clay tablets at Ashur (at the time of Tiglath-pileser I), and petering out in the fragmentary Neo-Babylonian laws dating from the 7th century BC.

The stela of Hammurabi must have been originally set up in some Babylonian population centre for the literate to read and know their rights. Some Elamite invader must have carried it off to Susa (perhaps *c.* 1200 BC), where it was found in 1901 and removed to the Louvre in Paris. The bulk of the stela contains the text of the code, partly erased on the obverse but restorable in some measure from clay-tablet versions of the same laws. The top depicts the king in a worshipful pose, receiving the laws from the sun god, Shamash. In reality Hammurabi—the sixth of 11 kings of the Old Babylonian or Amorite dynasty—was a practical codifier rather than a revelatory mediator of law. His code was an effort to fuse into a workable whole the ancient inheritance of Sumerian-based jurisprudence and the Semitic talion law (punishment according to the "eye-for-an-eye and tooth-for-a-tooth" principle) of the Akkadian superstratum. The result is not a model of economy or arrangement or logical organization, but the code of Hammurabi constitutes nevertheless the first great legal monument in the history of mankind. The later Assyrian laws show traces of further removal from the cradle of Sumerian civilization since they are both harsher and noticeably more primitive.

Ancient Egypt. Egypt attracted the special curiosity of the Greeks, and Herodotus (5th century BC) devoted an entire book to on-the-spot observations and fanciful tales about the land of the Nile. The lost *Aigyptiaka* (or *Aegyptiaca*) of Manetho (3rd century BC) contained the roster of 30 dynasties, which still underlies the chronology of ancient Egypt. Such classical writers as Strabo, Plutarch, and Pliny the Elder all dealt with various aspects of Egyptian antiquities.

Yet the fund of knowledge would be woefully skeletal and inaccurate without the explicit testimony of contemporary records from Egypt itself. The decipherment of the Egyptian writings gave the impetus to Egyptian epigraphy. The progress of excavations multiplied the corpora of texts, especially adding the papyrological dimension. In addition, cuneiform Akkadian on clay tablets was the international diplomatic medium of writing during the most brilliant phases of Egyptian history and is hence an integral part of the Egyptian epigraphic record.

The historically significant Egyptian epigraphic texts, apart from their external peculiarities, have likewise special traits relating to genres. There is little attempt at historiography and great fluctuation in bulk in the course of dynastic vicissitudes. They are mainly annalistic and thus firsthand accounts of pharaonic or other high-level deeds; but the peculiar features of stylization, stereotyping, and usurpation must frequently give the careful historian pause and sometimes debase the face value of the record. Written monuments became somewhat numerous during the 4th dynasty (*c.* 2575–*c.* 2465 BC), that of the pyramid builders Khufu, Khafre, and Menkaure. Notable are the fragmentary annals of Snefru, which already alluded to dealings with Asia, the chronic goal of Egyptian territorial ambitions. Historic records persisted under the following two dynasties, with particular articulateness in the reign of Pepi I, third king of the 6th dynasty (*c.* 2325–*c.* 2150 BC), then subsided until a modest reemergence in the (Theban) Middle Kingdom of the 12th dynasty (1938–*c.* 1756 BC). Another silence shrouded the period of the Hyksos kings (*c.* 1630–*c.* 1523 BC), broken only subsequently by such retrospective revulsion at the memory of barbarian domination as that in Queen Hatshepsut's (1479–58 BC) temple inscription at Istabl Antar in Middle Egypt. The golden age of historical recording began in the 15th century BC with the central rulers of the 18th dynasty, notably Thutmose III and Amenhotep II and III. Thutmose's annals on the walls of the temple of Karnak describe 20 years of ceaseless military activity in Asia, some 16 campaigns in all, and are supplemented by stelae from Armant in Upper Egypt and Gebel Barkal near the Fourth Cataract, as well as by lists of conquered lands at Karnak. Similar material continued in the reigns of Amenhotep II and III, in the latter's case importantly supplemented by the cuneiform correspondence with foreign powers (Mitanni, Arzawa,

etc.), which was subsequently stockpiled and archived by Ikhnaton in his transitory new capital, where it lay buried to await the modern excavators of Tell el-Amarna. Ikhnaton's religious preoccupations (he changed the official religion to the worship of the sun god Aton), and political apathy led to the loss of many of Egypt's Asian possessions. Records of Ikhnaton's short-lived son-in-law, Tutankhamen, at Thebes (1332–23 BC), make recantation and restoration for the heresy. Tutankhamen's successor, the warlord-pharaoh Horemheb, left boastful accounts of foreign conquest that sound suspiciously grandiose in relation to plausible reality.

In the 19th dynasty (1292–1190 BC) Seti I went to war against the Syrians, Hittites, and Libyans, letting the world know about it on the walls of Karnak. But in this respect he was no match for his long-lived son, Ramses II, who usurped the monuments of others and covered unprecedented amounts of wall space with his own real or inflated exploits. (The Battle of Kadesh against the Hittites in 1299 BC, which ended in a stalemate, was given lavish coverage as a triumph on temple walls at Karnak, Abydos, and Abu Simbel.) In the 20th dynasty (1190–1075 BC) occurred incursions of the "sea peoples," and the records of Ramses III detailed both the crisis and the increasing accumulation of wealth and power in the religious establishment. Subsequently, Egyptian history receded from the world scene, with "Libyan" and "Ethiopic" dynasties and a brief Saite renaissance of the 26th dynasty (664–525 BC), already under the Assyrian and Babylonian shadow, soon to be replaced by the Persian. The firsthand political records declined accordingly, although they remain of significance for local history down to the Ptolemaic era, a dynasty that ruled Egypt beginning in 304 BC, founded by Ptolemy I Soter, a general under Alexander the Great.

No law codes have been found in Egypt, presumably because codification was not practiced. There are, however, royal administrative and legal decrees granting privileges and immunities and also records of legal proceedings, especially of the Theban tomb-robbery trials during the 20th dynasty.

Other ancient Middle Eastern regions. Regions adjacent to the power centres of Mesopotamia, Egypt, and Iran were frequently mere political and administrative adjuncts, often obscure vassaldoms or adversaries without notable or attested written traditions. The Mitanni kingdom in northern Mesopotamia had some ephemeral big-power dealings with Egypt in the days of Amenhotep III, but its capital city is still lost in the sands, and thus its presently known epigraphic tradition is merely part of the correspondence in the Tell el-Amarna archives. The records of the Elamite kingdom with its capital at Susa were mostly ancillary to Mesopotamia in the 2nd millennium BC and to Iran later on. The region of Syria and what later came to be called Palestine was in the 2nd millennium the object of an extended tug-of-war between Egypt and the Hittite kingdom.

The Hittites were, in fact, the third great international power in the Near East during part of the 2nd millennium BC, and the epigraphic yield of their royal archives at Boğazköy in central Asia Minor matches or even surpasses in richness that of Mesopotamia and Egypt for the few centuries in question. The cuneiform records of the Hittites contain a tradition of unique royal political self-expression. These documents begin with the oldest known Hittite text, the inscription of the early ruler Anittas, detailing dynastic struggles of an obscure and possibly apocryphal past. From the founder of the Old Kingdom, the firmly historical Hattusilis I (Labarnas II), came an annalistic autobiography (excavated in 1957) and a "farewell address," or political testament, in Hittite as well as Akkadian versions. Subsequent events, including the capture of Babylon by Hattusilis' son, Mursilis I (*c.* 1590 BC), and the succeeding era of regicidal upheavals, are known from an edict of King Telipinus, who detailed them as he set about regulating the rights of royal succession. The subsequent founder of the Hittite Empire, Suppiluliumas I (*c.* 1350 BC), and his son Mursilis II left annals detailing their military and political deeds. Mursilis was a particularly prolific annalist and edited his father's annals as well.

The earliest Hittite texts

The great encounter with Ramses II at Kadesh in 1299 BC occurred in the reign of Mursilis' son, Muwatallis, and left an echo in the autobiography of his brother and successor, Hattusilis III. Hattusilis' autobiography is a tract of self-justification for a breach of the edict of Telipinus in deposing his nephew and predecessor Urhi-Teshub (Mursilis III).

Other Hittite documents inveigh against treasonable behaviour ("Indictment of Madduwattas") or contain detailed instructions for military, civil, and court officialdom. Hittite queens had prerogatives of independent high-level initiative, and examples of their correspondence with foreign potentates supplement the archives of their husbands. The most remarkable external political documents are numerous state treaties, sometimes between equals but more often covenants specifying protectorate or vassaldom status for subordinate states on the fringes of the kingdom. Equally notable is the Hittite Law Code, relatively enlightened and mild in the face of its contemporary counterparts in Mesopotamia. Altogether, the inscriptional documents are practically the exclusive source material for knowledge of the Hittites; not even the existence or location of their empire was surmised prior to the discovery of their archives.

After the collapse of the Hittite Empire (*c.* 1190 BC), significant records from Asia Minor ceased for many centuries, whereas local history in the Syro-Palestinian area was recorded in the inscriptions of petty dynasts increasingly under the shadow of Assyrian domination. The break with the past is evident in the writing systems (Hittite Hieroglyphs or West Semitic alphabet rather than cuneiform) and in the languages (Indo-European Anatolian, Canaanite, Aramaic). Into this category fall the stela of King Mesha of Moab (*c.* 830 BC) now in the Louvre, the Phoenician-Hieroglyphic Luvian bilingual inscription of Azitawadda of Adana (late 8th century BC), and those of the kings of Ya'diya-Sam'al. Contemporary cuneiform documents from the Urartu kingdom around Lake Van in eastern Anatolia are historically and culturally an offshoot of the history of 8th-century Assyria.

Ancient Iran. Epigraphically recorded history in ancient Persia began dramatically with the rise of the Achaemenid dynasty in the 6th century BC. Cyrus II the Great's conquest of Media, Lydia, and Babylonia, Cambyses' occupation of Egypt, and the incursions into Greece of the succeeding side branch of the family, beginning with Darius I, created in short order a world power destined for centre stage on the international scene for the following two centuries. The international character of the empire is reflected in the frequently trilingual royal inscriptions—with Akkadian and Elamite versions in traditional syllabic cuneiform, and the Old Persian text in its own simplified quasi-alphabetic system of wedge-shaped writing. The Achaemenids' time span ranges from Darius' great-grandfather, Ariaramnes, to their less glorious progeny, who were ultimately extinguished by Alexander the Great. The empire was centred in Persia; but a granite stela of Darius, found near the Suez Canal, recorded in Old Persian, Elamite, Akkadian, and hieroglyphic Egyptian the opening of a canal from the Red Sea to the Nile. The epigraphic material included rock surfaces, building walls, columns, doorways, cornices, statues, and doorknobs; bricks, plaques, plates, and tablets of clay, stone, gold, and silver; vases; weights; and seals. Almost all the longer texts were by Darius and Xerxes I; an important one of Xerxes was found on a stone tablet at Persepolis in 1967. The great Bīsitūn rock inscription of Darius runs to several hundred long lines in Old Persian alone, besides the Elamite and Akkadian versions. It is accompanied by 11 minor inscriptions, serving as keys to the sculpted scene of the panel, which shows Darius triumphant over the usurping impostor Gaumata and nine other rebels. The text is a self-statement of how Darius gained and consolidated his rule. It apparently had currency in the realm, apart from being tucked away on a sheer cliff wall, for a partial duplicate of the Akkadian version has been found on a dolerite (basalt) block from Babylon, and papyrus fragments from Elephantine have yielded scraps of an Aramaic edition. One peculiar documentary value of the text is that the

Rise of the Achaemenid dynasty

same period in Persian history is extensively covered in the Greek literary tradition by Herodotus, Ctesias, and others, and scholars can thus juxtapose Darius' own accounts with those of almost contemporary foreign historians. As an example, Darius stressed his role as saviour of the fatherland from the clutches of an upstart who pretended to be Smerdis, the brother of Darius' predecessor Cambyses. The latter had murdered Smerdis and was carrying on various outrages in Egypt when word came of the impostor's takeover back home. Darius stated that thereupon Cambyses "died his own death," meaning that it was a fatal matter without human interference, and that thus Darius' hands were clean in taking action against the impostor. According to Herodotus, Cambyses' legendary death involved a freak occurrence as he prepared to leave for home—an accidentally self-inflicted wound leading to gangrene. Both the meaning and intent of Darius' description are thus confirmed.

uxtapotion of arius' xts with erodotus' istorical ritings

From later Arsacid (Parthian) and Sāsānid periods of Iranian history, there are likewise royal inscriptions that shed light on their respective eras down to the Islāmic conquest in the 8th century AD, and new specimens are still being discovered.

Ancient India. India's past became anchored in historical time and separable from legend only with the establishment of firm synchronisms with outside data. One such link is the Seleucid embassy of Megasthenes to the Maurya king Candragupta (Greek Sandrokottos) at Pāṭaliputra (Greek Palimbothra) in Magadha (modern day Bihar). The Maurya dynasty was continued in the early 3rd century BC by Candragupta's son Bindusāra (Amitrochates in the Greek sources) and had extended its power over much of the subcontinent. But then the Greek sources fall silent, and Indic literary tradition supplies only the usual web of timeless legendry. At this point, however, epigraphy makes a unique contribution in the form of the first authentic and datable historical documents from India, the edicts of Bindusāra's son and successor Aśoka. As a matter of epigraphic fact, Aśoka ruled all of northern India and a large portion of the south, from Taxila and beyond to Mysore and Kaliṅga (coast of Orissa and Andhra Pradesh). His 14 rock edicts and seven pillar edicts in numerous versions and copies, plus separate minor texts, are scattered over this expanse—in the Prākrit language of his time and in the Brāhmī script, except for some northwestern examples of the Aramaic-inspired Kharoṣṭhī writing. Even a Greek-Aramaic bilingual version was found in 1958 near Kandahār in Afghanistan. Aśoka's edicts are proclamations and ordinances in a Buddhistic spirit, designed to impart good order, morality, and moderation by the Emperor's personal enjoining and example. Particularly notable is the 13th rock edict, which bares the ruler's pangs of conscience over the conquest of Kaliṅga eight years after his coronation, his continuing sorrow over the cruelties committed, and his pledge to substitute the victory of Buddhist religious law (*dhamma*) for all earthly conquest.

Aśoka's edicts would rate a mere historiographic footnote for their inconsequential transitoriness, were it not that in the same breath Aśoka supplied the very synchronisms that are the main key to ancient Indian chronology. Among his western neighbours he mentioned Aṃtiyoge (Antiochus II Theos of Syria), Tulamaye (Ptolemy II Philadelphus of Egypt), Aṇtekine (Antigonus II Gonatas of Macedonia), Maka (Magas of Cyrene), and Alikasudaro (Alexander of Epirus or Alexander of Corinth). The dates of these contemporaries circumscribe the time of Aśoka's reign; combined with the earlier Greek synchronisms, they afford a firm foundation for the correlation of Indic and Mediterranean events. The edicts of Aśoka are thus a prime example of the value of inscriptions for historiographic dating and constitute a fixed record unparallelled in ancient Indian tradition. Later periods of Indian history, such as those of the Indo-Scythian and Gupta rulers, are also represented in epigraphic documents of some historical value.

he main ey to ndian hronology

Ancient China. In China also, inscriptions are a means of separating chronological fact from historiographic legend. Nonepigraphic book composition on wood or bamboo strips had an early history in China, beginning in the later 2nd millennium BC; its scope was such that the Ch'in emperor Shih Huang-ti went down in history as a book burner in 213 BC. The San Tai, or three periods of early Chinese history (Hsia, *c.* 2205–1766 BC; Shang, *c.* 1766–1122 BC; Chou and Ch'in, *c.* 1122–206 BC), were long considered by Western scholars purely legendary down to the early Chou period, and the literary documents (such as the *Shu Ching* or "Classic of History") were dismissed as compilations consisting mostly of successive overlays of little historical value. But the historicity of written records from the later Shang era (*c.* 1400–1122 BC) is now apparent from the mass of inscribed archeological material found especially in northern Honan Province. These include, in particular, the so-called oracle bones (mostly tortoise shells and scapulae of animals), bearing incised records of royal divination. At the site of the Shang capital, Yin, were discovered inscribed vessels of bronze, bone, pottery, jade, and stone, probably ceremonial in nature and related to official ritual uses such as ancestor worship. The script is a mixture of pictograms, word signs, and phonograms. From the ensuing Chou era, bronze inscriptions of official provenance have likewise been found, especially records of royal largesse. Inscriptions from later periods form a steady but subsidiary source of information beside the larger, nonepigraphic written record.

Ancient Greece. The historically significant epigraphic record of classical Greece differs in many ways from most of those discussed above. Much of it is parallelled by a mature and independent tradition of professional literary historiography. Except for the preclassical Helladic (Mycenaean) period of the 2nd millennium BC (see below), there was no archival tradition, although the bulk of "monumental" records sometimes approximates the same purpose of massive preservation. There was no all-important power centre and no dominant rulership before Hellenistic and Roman times: thus the geographical scattering of records was extreme, although naturally with some focuses of emphasis such as Athens. Above all, there was continuity from the inception of literacy, with gradual but steady increase in bulk.

The special circumstances of the Greek epigraphic record

Epigraphically transmitted historiography in Greece is extremely scarce because the probing of past events has passed beyond the stage of dynastically centred and sheltered annalism; an example is the "Marmor Parium" (from the island of Paros and now at Oxford), which contains a chronographic rundown of traditional dates and events of Greek history. Rather than monolithic records of autocracy, there is history in the making by a plethora of tyrannical, oligarchic, or democratic microentities. Treaties of alliance and various other agreements between the multiple city-states—recorded on metal or stone and publicly displayed, or consecrated at such pan-Hellenic sanctuaries as Delphi or Olympia—form an important part of the epigraphic yield. Joint-citizenship covenants, decrees concerning the return of exiles, monetary agreements on coinage and debts, and tribute lists are typical examples. They are supplemented by the records of arbitration of interstate disputes, most often boundary matters, by third-party commissions. Thus, when around 240 BC a territorial disagreement arose between Epidaurus and Corinth, the Achaean League appointed a group of 151 Megarians as mediators, and their report survives. Further extensions of such "international" documentation are the *proxenia* decrees, which amount to letters of patent and resolutions of appreciation issued by one state to a citizen of another for service as *proxenos,* a kind of honorary consul looking after the interests of the other state's citizens. The extensive colonization efforts by the Greeks around the Mediterranean produced a further kind of political document—regulations governing conditions for emigration and return, citizenship rights of the colonists, and relations between the colony and the mother community. Not all historically meaningful international records are of the monumental type. Greek mercenaries of Pharaoh Psamtik II (ruled 594–589 BC) left their scrawlings on the legs of a colossal statue at Abu Simbel on the upper Nile, proving by their names and dialect that they came from Rhodes and Ionia and were far abroad on foreign adventure.

Internal documents of the various Greek states include numerous records of decrees and ordinances, both administrative and legislative. Stereotyped Athenian ones are complemented by variant forms in other localities; most contain a preamble setting forth the date and the officialdom in charge, the circumstances occasioning the action, the decision itself, means and sanctions for its enforcement, and sometimes instructions for providing and affixing the very physical record that has been preserved. Sometimes they amount to formal laws, such as those directed against extravagances in funeral practices.

Financial data of the states were minutely and permanently inscribed on stone, and the accounts thus displayed recorded in detail the receipts, expenditures, and balances of public funds. Very specific reports cover projects of public construction, including both technical and budgetal details, allowing sometimes the integral modern reconstruction of the buildings from the reports alone. The records of the Erechtheum and the Parthenon at Athens are well preserved, as are inventories of military expenditures, especially those of the Athenian navy. Knowledge of the ephebic system at Athens, a paramilitary youth organization, is in the main based on epigraphic material.

The laws of Gortyn

The only law code in the Greek epigraphic tradition is the laws of Gortyn in central Crete, inscribed on the slabs of a circular wall which, if completely preserved, would have been nearly 100 feet (30 metres) in diameter. The 12 columns of text, each on four layers of stone and some five feet (1.5 metres) high, are about 30 feet (nine metres) in sideways length and contain more than 600 lines of text, being the longest Greek epigraphic monument; parts of some columns of further text survive, the so-called Second Code. The probable date of this inscription is the first half of the 5th century BC. The code deals with such matters as disputed ownership of slaves, rape and adultery, rights of a wife upon divorce or death of husband, disposition of children born after divorce, inheritance, sale and mortgaging of property, ransom, children of mixed marriages, and adoption. While self-contained, it evidently does not represent the entirety of laws; curiously, it stresses those areas of civil law (inheritance, adoption) that are notably lacking in the Hittite Law Code. The uniqueness of this code in the Greek world points up the relative isolation and marginality of the Cretan tradition, with tendencies to codification more reminiscent of Anatolia and the Near East generally.

Ancient Rome. While partly overlapping Greek inscriptions in time and type, those of Rome nevertheless present distinct peculiarities. There is a high measure of standardization in kind and style, despite lingering local traditions in more remote areas. Extensive and excessive use was made of initials and abbreviations, to the point of serious impediments to comprehension; lists of such abbreviations are standard adjuncts to modern handbooks on Latin epigraphy. Stone and bronze were standard material, but there was more use made of bricks, tiles, and terra-cotta, and practices of stamping and signing such matter are of help in identification and dating.

Literary and epigraphic records of early republican Rome are scant and fragmentary. Latin was at the time still largely confined to Rome proper, with Oscan, Etruscan, and colonial Greek spoken and written in much of Italy. With the arrival of extended political power there was little early literacy to fall back on, and historiographic attempts at retrospection ended in epicized myth and legendry (*e.g.*, in Livy). The Greeks on the southern coastal fringe had little truck with the hinterlands of early times. The Etruscan impact on Rome is evident, but shortcomings in discovering epigraphic records of Etruscan city sites (as opposed to necropolises) and in understanding the Etruscan language, limit the historical data derivable from Etruscology. The potential for such illumination is seen from the discovery of gold tablets at Pyrgi in 1964 that contain a dedication in Etruscan and Phoenician by the Etruscan king of Caere, Thefarie Velianas, to the syncretized goddess Uni (Juno) Astarte. Datable to around 500 BC, the text shows Etruscans ruling in the outskirts of Rome, with enough Phoenician or Punic (Carthaginian) maritime presence to warrant symbiotic and syncretistic bilingualism. The vital historical import of such attestations, pieced together with later Greek and Roman historiographic data, is patently manifest.

Texts from republican Rome

No historically important epigraphic Latin text from republican Rome antedates the 2nd century BC. The marble Columna Rostrata—found in Rome in 1565 and now at the Palazzo dei Conservatori on the Capitoline Hill—records a naval victory of Duilius (consul in 260 BC) over the Carthaginians; but the inscription, replete with fake archaism, dates from a restoration effort in early imperial days. The *fasti consulares* and similar lists afford a summary sequence of consulates, magistratures, and triumphs. The one truly significant epigraphic historical text is the "Res gestae divi Augusti," an autobiographical record of Augustus' rule, which was exhibited in many places but is best known as the Monumentum Ancyranum, from the bilingual (Latin and Greek) version carved on the walls of the Temple of Rome and Augustus at Ankara (Turkey).

By the time the epigraphic record became abundant, Rome's domination was secure and the political documentation was one of imperial outflow and of local sycophancy. Treaties of republican Rome with foreign powers survive merely in the works of literary historians. Among "internal" documents from republican days are several epigraphic texts of significance: the Senatusconsultum de Bacchanalibus, on a bronze tablet found in 1640 in Bruttium (the "toe" of Italy) and now in Vienna, is a consular edict on Senate authority, regulating Dionysiac outbursts in Italy in 186 BC; pieces of the laws Lex Acilia Repetundarum (123 BC) and Lex Agraria (111 BC) were found in the 16th century on opposite sides of what was once a large bronze tablet; the local laws of the town of Bantia (on the borderlands of Lucania and Apulia in southern Italy) are inscribed on a fragmentary bronze tablet found in 1790 (now in Naples), with a Latin-language text on one side and the longest known Oscan inscription on the other, both datable to the late 2nd century BC; parts of the Lex Cornelia de Viginti Quaestoribus (81 BC) are preserved on a large bronze tablet found at Rome; Julius Caesar's Lex Julia Municipalis of 45 BC was found near Heraclea in Lucania. On the whole, however, the transmission of Roman law, from the earliest fragments to the mature codifications, is nonepigraphic. In later times the flood of administrative decrees increases with the growth of centralized autocracy. Typically Roman epigraphic material of imperial date comprises further building inscriptions, military records, and honorific texts.

The oldest Turkic monuments

The Turkic peoples. The oldest monuments of Turkic languages—inscribed on stones, and datable to the early 8th century AD—were discovered in the late 19th century in southern Siberia around the Yenisey River and in northern Mongolia near the capital of Urga (modern Ulaanbaatar). Deciphered in 1893 by the Danish scholar Vilhelm Thomsen, they provide valuable insights into the history of Central Asia around the 7th century AD. These records of the Turk dynasty (Chinese T'u-chüeh) comprise especially texts found at Kosho-Tsaidam on the Orhon (Orkhon) Gol (river), including also Chinese text. These texts throw light on the nomadic culture of the tribal empire controlled by the Turk dynasty, including shamanism, calendar, customs, and social structure, with strong Chinese influence detectable in the latter. After the decline of the Turk people (*c.* 745), their successors, the Uighurs, perpetuated for a time the same kind of monumental dynastic epigraphy, the writing system of which is an offshoot of the Aramaic alphabet, presumably mediated by the Iranian-speaking Sogdians of Central Asia. Gradually, however, new scripts took over (especially the so-called Uighur alphabet, of Syriac origin, which was further transmitted to the Mongols and the Manchus) and inscriptional monuments gave way to manuscript records such as those found in Chinese Turkistan (Turfan) in the late 19th century (along with texts in Sanskrit, Sogdian, Tocharian, and other Indo-European idioms), attesting to a coexistence of Buddhist, Manichaean, and Nestorian Christian religious communities. The later Turkish peoples, including the Anatolian Seljuqs and Ottomans, had an Islāmic book-tradition, to which the inscriptional record is merely incidental.

Northern Europe. The advent of writing was slow north of the Alps; it came either from direct expansionary exportation by Greek coastal colonies and the Roman Empire, as in Gaul and the Iberian Peninsula, or indirect inspiration from the same quarter, as in writing in the Irish and British ogham alphabet and the Germanic runes.

Celt-Iberian inscriptions from Spain and Celtic ones from Gaul and Ireland are scarce, mostly brief, and notably devoid of usable historical information, apart from their mere monumental existence and linguistic and onomastic (pertaining to names) content. Occasional items such as the fragmentary Gaulish Calendar of Coligny afford insights into local cultural practices, apart from an overwhelming trend to romanization.

The runic alphabet—a Germanic alphabet, originally of 24 letters, also called futhark—and its offshoots (the Scandinavian, especially Danish, 16-letter variety from the 9th century AD; and Anglo-Saxon versions, from the 3rd to the 10th centuries AD, also called futhorc) are probably of "North Etruscan" or "Sub-Alpine" Italic inspiration, datable to around 200 BC. The "North Italic" letters of the Germanic text *harixasti teiva,* "to the god Harigast," on a helmet from Negau (southern Austria) are probably from that time of transmission. Runic inscriptions from the era of migrations, ranging from eastern France through Germany up to Denmark and eastward via Poland to Romania, are supplemented by the later, richer yield from England and Scandinavia. Native Anglo-Saxon runic epigraphy, mostly in Northumbria, Mercia, and Kent, petered out around the 10th century, whereas the Scandinavian tradition (including its enclaves on British soil) endured for several more centuries. Sweden has some 3,000 runic monuments; Norway and Denmark, perhaps 400 each; while Iceland has remarkably few, apparently in inverse proportion to the literary flowering in that colonial outpost. The Vikings left their runic calling cards in far-flung places, including those in the Greek port of Piraeus, on the Black Sea coast, in Varangian Russia, in Scotland, Ireland, and the Isle of Man, and the Orkneys, Hebrides, and Shetland islands; Greenland has its share, but alleged specimens from North America, notably the Kensington Stone from Minnesota, telling of the westward trek of an exploration party from Vinland, are invariably crude latter-day forgeries.

Runic monuments

The purposes of runic inscriptions were usually either dedicatory or commemorative, sometimes magic, and frequently sepulchral. The longest, that from Rök in Sweden (725 runes), seems to contain a catalog of epic deeds, possibly those of the Ostrogoth king Theoderic. The prime historical value of runic epigraphs is usually what and where they are, rather than what they depict or record.

Inscriptions as social and cultural records. In the preceding section, inscriptions were evaluated as sources for the presence and migrations of peoples, the existence and chronology of political states, their dynastic histories, foreign relations, internal governance, legal institutions, and official acts. In this section, epigraphy is surveyed for information about how past civilizations lived; their religious beliefs and practices; their business, financial, legal, and social relations; and what shape their aspirations assumed in terms of verbal creativity. The subdivision of civilizations surveyed differs somewhat arbitrarily from the earlier section by the omission of certain areas and the inclusion of Crete and Mycenaean Greece. In fact, Indic and Chinese epigraphic matter discussed above could just as well have fitted the "religious" slot, but its royal character and chronological importance for official history dictated otherwise. Conversely, the Cretan and Mycenaean tablets are purely economic inventories, but they might possibly have been included with history above for the very important historical fact that they prove the ruling presence of Greeks at Knossos during the 2nd millennium BC. The varying degree of importance of epigraphic material in various cultures persists: in Mesopotamia and the ancient Near East its dominance was nearly total; in Egypt it combined with the papyrological dimension; in Crete it was merely a flash in a prehistoric darkness; while in ancient Greece and Rome, it was a supplementary concomitant of the nonepigraphic literary tradition.

Ancient Mesopotamia. Ample specimens of Akkadian-language clay-tablet epistolography have been found at several sites, notably Tell el-Amarna in Egypt and Tell al-Ḥarīrī on the middle Euphrates (the ancient Mari of *c.* 1700 BC). The Amarna letters, about 400 of them, were composed in corrupt Akkadian by Canaanite scribes in Syria and Palestine and were largely official in character. The Mari letters, some 5,000 in number, are more illustrative of normal day-to-day written communication in a Mesopotamian milieu proper.

The Amarna and Mari letters

Another aspect of everyday life in ancient Mesopotamia is amply illustrated by thousands of clay tablets of a practical legal nature, as distinct from the formal laws. These archivally preserved records from various periods use Sumerian, Old Akkadian, Babylonian, and Assyrian alike. They detail law suits, court decisions, marriage contracts, divorce proceedings and settlements, sale adoptions (fictitious acts circumventing prohibitions against land sales outside the family), loan agreements, tax receipts, and much else. Commercial inventories, such as those of the Old Assyrian merchant colony at *Karum Kanes* in central Asia Minor (20th century BC), complete the picture.

Due to the religious sanction of law, legal records were often stockpiled in temple archives. These latter are also the source of more directly cultic texts, such as descriptions of rituals, which come under such headings as "Temple Program for the New Year's Festivals at Babylon," "Ritual to be Followed by the Kalū (priest) when Covering the Temple Kettle-Drum," "Ritual for the Repair of a Temple," and "Program of the Pageant of the Statue of the God Anu at Uruk." Prayers, lamentations, and hymns in both Sumerian and Akkadian are extant, addressed to deities such as the goddess Ishtar, the moon god Sin, the sun god Shamash, the great triad Anu, Enlil, and Ea, and the Babylonian patron god Marduk. The Sumerian "Lament for the Destruction of Ur" bemoans the city's fall to Elamites and Subarians. Often the king himself is the spokesman in the text. Wisdom literature, such as proverbs and fables (*e.g.,* "Dispute between the Date Palm and the Tamarisk"), poetic meditations, oracles, divination records, omens, and prophecies are further examples of Mesopotamian genres that only epigraphy has preserved.

Sumerian and Akkadian narrative literature is likewise of wholly inscriptional transmission. It contains man's earliest preserved literary creations in the Sumerian sequence, especially the texts from tablets found at Nippur. These include the "Paradise myth" of the god Enki and the goddess Ninhursag in the pure, clean, and bright land of Dilmun; the story of Dumuzi and Enkimdu (the petulant shepherd god versus the peace-loving farmer god, inversely reminiscent of the Cain–Abel antagonism in Genesis but not culminating in murder); "The Deluge" with its Noah-hero Ziusudra; "Inanna's Descent to the Nether World," which prefigures the later Akkadian "Ishtar's Descent"; and the lays of Gilgamesh, which show the Sumerian traditions that were later partly organized and transformed into the Akkadian epic. The latter include "Gilgamesh and Agga of Kish," a story of confrontation between early Sumerian city-states; "Gilgamesh and the Land of the Living"; and "The Death of Gilgamesh," with its haunting parallelistic refrain "he lies, he rises not." The Akkadian *Epic of Gilgamesh* is divided into 12 tablets, the longest of which is more than 300 lines; this "Flood Tablet" (the 11th) is virtually intact and comes, like almost all Assyrian-language Gilgamesh texts, from the library of Ashurbanipal at Nineveh (7th century BC). From the 2nd millennium there are fragments of a Hittite version from Boğazköy, as well as minor traces of a Hurrian translation. Old Babylonian correspondences to tablets 1–3 and 10 are found on a tablet from Sippar (*c.* 1800 BC). The 12th tablet is a literal translation from Sumerian, whereas the rest amounts to a self-contained Akkadian epic original, based on Sumerian motifs but with a thrust of its own. The most complete reconstruction involves a combination of Assyrian, Old Babylonian, and Hittite versions.

The Gilgamesh epic

The other famous Mesopotamian epic, *Enuma elish,* "When on high," details the story of cosmic creation and of how Marduk became the great god of Babylon; it had

more immediate cultic attachments because its recitation formed part of the New Year festival.

Further Akkadian literary creation is attested in the epic of Atrahasis, a tale of mankind's punishment through pestilence and flood, preserved in fragmentary Old Babylonian and Assyrian versions. The story of Adapa, found in parts in the Tell el-Amarna archives and the library of Ashurbanipal, is similar to Gilgamesh's quest for immortality. The myth of Zu deals with the theft of the tables of fate and the usurpation of almightiness by the bird god Zu. The legend of Etana, a namesake of the shepherd-king who ascended to heaven in the mythical postdiluvian Sumerian dynasty of Kish, recounts in its Old Babylonian and Assyrian recensions the heavenly flight of Etana on the wings of an eagle in order to acquire the magic birth plant that would cure his childlessness. Death-oriented themes appear in the tale of Ishtar's descent, in the story of Nergal and Ereshkigal, and in various netherworld texts associated with the Tammuz myth and liturgy.

The accidental preservation of Egyptian texts

Ancient Egypt. The multitude of incidental day-to-day written matter on potsherds and papyrus, preserved by climatological quirks in Egypt, affords insights into the normal living patterns of ordinary people during the pharaonic, the Hellenistic (Ptolemaic), and the Roman periods. More formal records of similar provenance may be described as legal in the widest sense, comprising such objects as land deeds, cadastral inventories (tax surveys), wills, adoption decrees, and trial transcripts. Thus many technicalities of landholding in the Nile Delta during certain ancient periods survive, unwarranted by any normal expectancies of epigraphic preservation elsewhere; only the happenstance of the Creto-Mycenaean records (see below *Crete and Mycenaean Greece*) has accidentally transmitted something analogous.

Ritualistic texts are equally abundant, in both monumental and papyrological transmission. Notable among offertory liturgy was the mortuary service to the dead, acted out as a kind of ritual make-believe in which the offering was referred to as the eye of Horus (the sacrificer) being given up on behalf of his father, Osiris (the departed). Ritual drama in general is amply attested in Egyptian religious practice; day-to-day ceremonial in the cult of gods or god-kings (pharaohs) was recorded in the minutest detail. Magic texts include incantations and charms against such dangerous creatures as snakes, scorpions, crocodiles, and lions. Even the buried dead were deemed in need of such protection, and hence the preserved samples are largely of mortuary provenance. An important category of epigraphs is constituted by curses, imprecations, and threats of divine vengeance against enemies, evildoers, and tomb violators in particular. Dream interpretation and oracular inquiry are represented on both monuments and papyri.

Of a more secular nature and on the verge of true literature are moral and didactic writings, particularly during the early Middle Kingdom (began 1938 BC), when a profound social and spiritual crisis seems to have gripped Egypt. Of such kind are "The Admonitions of Ipuwer" (a denunciation of current sin and evil in Hebrew "prophetic" manner), the "Colloquy of a Prospective Suicide with His Own Soul," and especially copious compendia of conventional wisdom, frequently formulated as instructions from a royal or noble father to his son. Some parts offer remarkable parallels to the Hebrew book of Proverbs.

On the same borderline to real literature stand mythical tales, including in particular the myths of creation by the god Atum of Heliopolis and the Memphite patron deity Ptah, whose eminence is reminiscent of the cultic fortunes of Marduk. The Memphite cosmogony, of 1st dynasty provenance, is an abstract one, with Ptah's creation conceived with his "heart" (mind) and "tongue" (word) rather than the usual physical detail. Other mythic tales concern the perils from the netherworld demon Apepi faced by the sun god Re during his nightly journey in his solar boat; the rivalries between Horus and Seth for the succession of Osiris (a patent allegory of the tensions between Lower and Upper Egypt); mortuary texts particularly concerned with the Osiris myth and the prospect of judgment (*e.g.*, the "Protestations of Innocence"), attested both monumentally and on papyri (the Book of the Dead); and pieces designed to bolster the supremacy of specific sites and deities, particularly Amon-Re at Thebes.

Out of the latter tendency emerged hymnic poetry during the 18th and 19th dynasties, triggered by an imperial trend toward unicentrism in government and universal syncretism in religion. The great hymns to Amon-Re were matched during Ikhnaton's "Amarna revolution" by his great Aton-hymn, uncannily reminiscent of the 104th Psalm of the Old Testament. Victory hymns and prayers are further examples of poetic genre, as are purely secular pieces, including several collections of love songs from *c.* 1300 to 1100 BC, songs of farm workers and herdsmen, and feast songs for entertainment.

In nonreligious literary prose there is preserved the story of Sinuhe, an account of the life of a Middle Kingdom official, his prosperous peregrination in Asia but continued yearning for the homeland, relieved at last by a chance to join the pharaonic court. Papyri and ostraca tell this story in a long succession of attestations through most of the 2nd millennium BC. Of later date (*c.* 1100 BC) is the "picaresque" narrative of the hieratic-script "Moscow papyrus," detailing the ill-financed mission of a certain Wen-Amon from Karnak to Byblos to buy timber for a ceremonial barge of Amon.

The Boğazköy archives

The Hittite Empire. The Boğazköy archives are the unique central storehouse of Hittite records for the duration of that empire (only minor additions have been found elsewhere, such as letters at Tell el-Amarna and Alalakh). The past of other cultures was known from external and nonepigraphic sources before the explorational and excavational discoveries of inscriptions; with the Hittites, however, the record is in one place and *in toto.*

Much may be learned of everyday life and social relations from legal texts, including the laws themselves. Such matters as family relations, property rights, land tenure, and commodity prices are dealt with in great and sometimes pedantic or amusing detail. More can be learned from gift deeds of land, cadastral lists, and trial records, the latter containing largely depositions of respondents and witnesses in civil suits. Unique in their kind are treatises on the training of racehorses. Astrological omen texts, divination records (hepatoscopy, auspicy, lottery oracles), and dream interpretations abound, patterned largely on Mesopotamian models; rituals of all kinds take up a disproportionate amount of space. Temple foundations and procedures, festival ceremonies, epidemics, impotence, family quarrels, countermagic, royal welfare, birth, insomnia, and the manipulation of deities are among the topics of Hittite ritual; in a class by itself is a funeral ritual for royalty, reminiscent of Homeric practices. Not only the Hittite language but Hattic, Hurrian, Luwian, and Palaic are used in rituals. Another large category is made up of hymns and prayers, most notably those of King Mursilis II, in particular his "plague prayers," in which he confesses sins and begs the storm god for mercy. Other notable examples are the prayer of Muwatallis to the sun god, the vows of Queen Puduhepa to the sun goddess, and the prayer of Kantuzilis with its intimations of mortality. Native Anatolian mythical texts, such as those of the slaying of the dragon Illuyankas and of the disappearance of the god Telipinus, were part of ritual recitation. Other mythical matter points up the crossroads character of Hittite civilization, being made up of Akkadian, Hurrian, and Canaanite themes. Most of the Hittite epigraphic corpus has some connection with religion.

Other countries of the ancient Middle East. Away from the big power centres some quite important sites remain to be identified, such as the Mitannian capital of Wassukkani. Others (*e.g.*, the fortress-town of Carchemish on the upper Euphrates, for some time a Hittite dependency) have yielded notable data, especially royal inscriptions in Hittite hieroglyphs. In particular the French excavations at Ras Shamra on the Syrian coast since 1929 have uncovered the inscriptional and other remains of the small but strategic city-state of Ugarit, which flourished in the 15th–13th centuries BC. Its own vowelless cuneiform alphabet, a radical departure from the Mesopotamian syllabary prototype, denotes an archaic Semitic dialect closely akin to Canaanite. The written documentation of this crossroads

The city-state of Ugarit

community, in a variety of scripts and languages, provides information on the history of the area where Hittite and Egyptian power politics collided. But its most significant epigraphic products have been poetic epics of mythical and heroic scope, which go far to elucidate religious traditions otherwise known mainly from Old Testament bias. The great cycle of Baal and Anath brings to life the pantheon to which belonged also El, Asherah (the Astarte of the Phoenicians), Kothar ("Deft," the craftsman god), and Yamm (the sea god). The tale of Aqhat is also on the borderline of myth, while that of King Keret (or Kret) takes place in a saga world somewhat reminiscent of the Greek Homeric tradition.

Scattered dedicatory inscriptions and papyrus texts shed some light on the lives and practices of the Aramaean-speaking populations during the 1st millennium BC. From Palestine there are the Hebrew ostraca of Samaria, datable to the reign of Jeroboam II of Israel (8th century BC), which record names, families, and administrative and religious practices. Of equal significance are the ostraca of Lachish in southern Palestine, which probably immediately preceded the Chaldean onslaught of 589 BC. Phoenician texts are scattered around the Mediterranean, and bear witness to an extensive and protracted maritime supremacy.

The Old Persian Achaemenid inscriptions (see above *Ancient Iran*) are of some value in assessing the state religion of the time, which seems to have been a rather anemic form of official Zoroastrianism. Later monuments from the wider Iranian area help map its complicated religious history, such as the great inscription of the Kuṣāṇa king Kaniṣka, found at Surkh-Kotal in Afghanistan in 1957 and attesting to the Iranian language and Mithraic cults of ancient Bactria in the 2nd century AD. The Sāsānian religious tradition of the 3rd–7th centuries, with its rigidified formal Zoroastrianism, is mainly of nonepigraphic attestation.

Crete and Mycenaean Greece. The decipherment of the latest and most copiously attested of the Minoan linear scripts, the so-called Linear B, by British cryptologist Michael Ventris in 1953, is a major example of the dramatic impact epigraphic discovery can have on the most varied antiquarian disciplines. It supplied incontrovertible proof that the Mycenaeans on the Greek mainland during the 2nd millennium BC were Greek in language and likewise that Knossos in Crete was a Greek-speaking stronghold at the time of its final destruction (*c.* 1400 BC). The records from Knossos, Pylos in Messenia, and Mycenae in the Argolid are exclusively perishable inventories on clay tablets, kept in royal palaces or emporia and apparently meant to be discarded or reused annually; thus they record data from the very last year of any given establishment, just before its final destruction in the fire that also accidentally ensured their preservation by baking the clay.

The Linear B script

The contents are brief, concise, practical, and quantifying in character. Nevertheless, they reveal an important amount of detail about a civilization that is otherwise accessible only via the testimony of archaeology, iconography, and the poetic memory of early classical Greece across several "dark" centuries. Apart from the philologically invaluable recording of preclassical linguistic forms and proper names, there is the cryptologic evidence of the syllabic writing system, which may yet help bring about a cogent decipherment and linguistic identification of the earlier forms of Cretan writing.

Something of bureaucratic accounting methods can be learned from the system of the tablets. Contents yield lists of personnel; of livestock and agricultural produce; of textiles, vessels, and furniture; of metals and military matériel (swords, corselets, chariots, etc.); as well as records of landholding and of temporal and cultic tribute. The cadastral inventories are significant for the light they shed on the system of land distribution as compared with classical Greece and especially with contemporary practices in adjacent culture areas, such as those recorded in the Hittite Law Code. The religious tributes are revelatory in proving what bold scholarship had merely surmised—that the bulk of the classical Greek deities existed in name and kind already in Mycenaean days. Something of the social order and the system of government can also be incidentally learned.

Classical Greece. Inscriptions in their variety and profusion are an important means of gauging everyday life in classical Greece, especially such more formal aspects as were deserving of recording. This qualification, however, does not prevent some obscene pederastic rock-carvings on the island of Thera from being the earliest epigraphic texts from the classical period (*c.* 8th century BC), in a form of the alphabet still strongly marked by its Phoenician source.

Earliest classical epigraphic texts

In a formal vein there is ample documentation of a business, financial, and legal type, such as regulations for building projects, contracts for medical services (*e.g.*, the Cypriot bronze tablet of Idalium, in a peculiar syllabaric script reminiscent of the Cretan and Cypro-Minoan types, used to write both Greek and the uninterpreted Eteocypriot language), collection agreements for defaulted notes, manumission decrees, records of bank deposits (frequently in a temple), and reports of land commissions (especially the Heraclean Tables from southern Italy).

Even in these formal documents it is impossible to keep out reference to religious establishments. Religion was, in fact, big business in classical Greece, and overlap in the records is frequent. Temples were often of the order of corporations, with extensive holdings of real estate and banking operations. The epigraphs of their transactions are coupled with minute records of dedicatory offerings, priest lists, administrative regulations, and ritualistic and liturgical instructions. So much of social life was centred around institutionalized religion that most of its documentation may be included under the latter heading. Dedicatory texts in the widest sense are common, often in verse, such as that by Nikandre of Naxos on a 7th-century-BC statue of Artemis at Delos, a kind of propitiatory offering to the maiden goddess on the occasion of marriage. Similar hexametric or elegiac texts are found as epitaphs, especially for those who perished in war or at sea.

Tomb inscriptions generally are the most numerous of all but tend to be laconic in style and lacking even in purely statistical data. Those from ancillary areas of the Greek-speaking world, such as the non-Greek ones from Lycia, Lydia, and Phrygia in Asia Minor, are sometimes more revealing to the extent that they are interpreted. Thus the Lycian stela of Xanthus, a tomb monument from the 5th century BC, runs to many hundreds of words, including a dozen lines of Greek. The peculiar Lycian system of matrilinear descent is clearly evident in the texts. The Lydian ones include a marble stela from the necropolis of Sardis with a Lydian-Aramaic bilingual epitaph that typically calls down the curse of Artemis of Ephesus on potential violators. Such imprecations are also standard in Phrygian sepulchral epigraphs. Some kinds of death-related inscriptions were intentionally buried, such as the curses that consigned the object to the infernal avengers. Another type is seen in the Orphic tablets—texts on thin gold found interred with the remains of devotees of the Orphic salvation religion near Sybaris in southern Italy, near Rome, and at Eleutherna in Crete. They are apparently meant as instructions to the deceased—exhortations to the soul on its progress in the beyond, enjoining the shunning of the spring of Lethe ("Oblivion") and the drinking from that of Mnemosyne ("Memory"), thus securing the end of transmigration of the soul and the attainment of perpetual higher consciousness.

Tomb monuments

The Orphic texts provide the popular parallel to the Platonic myth of Er in the 10th book of *The Republic,* where Plato somewhat recast Lethe and Mnemosyne into Ameles ("Unmindfulness") and Anamnesis ("Recollection"). Thus, epigraphy not merely supplements the literary record of Greek civilization but also complements it in important aspects. If papyrology is included, important additions to the preserved portion of classical Greek literature have come from Egypt, especially in the papyri from Oxyrhynchus; these include such otherwise lost treatises as Aristotle's *Constitution of Athens* and entire dramatic works (*e.g.*, the *Dyscolus* of Menander).

Ancient Rome. The frequency of Roman inscriptions increased dramatically in direct proportion to the rise of

Roman power; but that same rise brought centralization, stereotyping, and a certain sterility of the more formal parts of the epigraphic record. Much of the bulk is official, unidirectional, and from the top outward. The Greek-speaking parts of the empire continued in many of their ancient ways, and other annexed regions (such as Spain or Gaul) developed their own locally coloured practices. The spread of a variety of exotic religious cults all over the empire, among them Mithraism and Christianity, added a kind of epigraphic underground initially devoid of official sanction and largely unmatched by alternative avenues of preserved written transmission. Popular epigraphy, including such matter as graffiti at Pompeii and other Vulgar Latin inscriptions, provides further counterpoise to the official stereotypes.

From early republican days the Roman written record is very spotty. The earliest text of any length, the Forum inscription from the early 5th century BC, seems to refer to an augural rite (referred to by Cicero and Festus as the *juge auspicium*) which enjoins the immediate unyoking of beasts of burden that produce excrement. A 4th-century vase (the "Quirinal vase") is apparently a wedding present with an inscribed guarantee by the bride's guardian to use his good offices to further marital concord. Later texts of religious import include the "Song of the Arval Brethren" on a marble tablet found in the Vatican in 1778 (a chant largely reduced to gibberish by age-old ritualistic repetition), records of various other cultic "colleges" (*augures; fetiales,* or priestly diplomatic representatives; *salii,* or priests of Mars and Quirinus et al.), and outside of the Latin-speech area especially the seven Umbrian-language bronze tables found at Iguvium (modern Gubbio) in 1444, recording in more than 4,000 words the ritualistic details of a brotherhood (the Fratres Atiedii) that flourished in republican days. Official religion is further seen in many dedications, such as one found on Caelian Hill in Rome in the 18th century, which records the offering of a temple and statue to Hercules by Lucius Mummius in 146 BC in fulfillment of a vow he had made during his conquest of Greece and sack of Corinth. "Unofficial" cults, especially those of Mithra and Jupiter Dolichenus, can be traced and mapped throughout the empire chiefly with the help of dedicatory epigraphs. Sepulchral inscriptions are recorded in overwhelming numbers, not only in Latin but notably also in Etruscan; among the latter is the mummy wrap, a lengthy funeral liturgy inscribed on an apparently Etruscan mummy entombed in Egypt (now in Zagreb). The earliest notable Roman examples are the epitaphs in Saturnian verse on the sarcophagi of the Scipios from the 3rd century BC.

There are many texts of deep sentiment and poetic feeling. Yet the bulk itself is the main source of information because it affords what may be called demographic statistics, including population distribution, occupations, public health, and longevity. The frequent inclusion of the *cursus honorum,* or career summary of the departed, has similar informational value. Tombs were usually placed under the divine tutelage of the powers of the beyond. Imprecations against violation were common, as were buried curse tablets in Latin and in Oscan (the *defixiones*). Many of the unofficial inscriptions exhibit substandard degrees of literacy and colloquial features, being thus an important adjunct to the study of the totality of classical Latin and its subsequent developments.

The use of inscriptions. *The dating of historical events.* Inscriptions are important specimens for chronology because they are often physical objects contemporary in execution with their contents. The dating of the inscription itself frequently yields a trustworthy chronology of its message: a victory stela records something freshly deserving of celebration; an epitaph implies a recent death. Exceptions do exist, which record more or less remote events at a conscious historical remove; archival specimens, for example, and secondhand copies generally lack the contemporaneity of other inscriptions. On the whole, however, external dating is crucial and may be achieved in several ways. Excavated monuments can be chronologized by their archeological context, including stratigraphic analysis and radiocarbon dating of any adjacent remains of organic matter. The shape of the monument may permit stylistic and iconographic determination. The type and variety of script used, and especially the style of writing, often allow paleographic dating. Thus, the relative age of Hittite texts can be determined by spotting the typical "Old Hittite ductus" of the more ancient period, and the various "scribal hands" of the Linear B tablets have been differentiated with extreme subtlety. Sometimes a radical reform, such as the official adoption by Athens of the Ionic alphabet in 403 BC (replacing the local Attic variety), provides a chronological watershed. Internal evidence of the inscription may yield its own kind of dating, either by synchronism with otherwise known facts or events or in true calendaric fashion. The year is frequently indicated by a king's reign or the tenure of a magistrate. Such stable counting of time as, for example, from Rome's legendary founding in 753 BC, or from creation (5509 BC, according to Christian dating) is rare in inscriptions; it became more of a historian's device during the classical and postclassical periods. In smaller communities, however, especially in Asia Minor, analogous local departures were used (legendary or historical foundation dates or other epoch-making events), with confusing results for latter-day chronologists.

Contemporaneity of epigraphy

The use of inscriptions for the dating of historical events is most pervasive when the historical tradition itself is "timeless," as in ancient India; the entire Indian chronology comes to be anchored around the Aśokan inscriptions. Inscriptions also permit a check on the veracity of ancient historians such as Herodotus (dubbed both "father of history" and "father of lies"), as in the case of the Bīsitūn inscription of Darius. Equally dramatically, the Linear B tablets prove at one stroke that the Greeks were ensconced at Knossos in the 2nd millennium BC and that the bulk of the "Olympian" religion was already theirs at the time. Most significant of all, nothing would be known of the great Hittite Empire during the 2nd millennium BC, were it not for the discovery of its inscriptional archives.

The decipherment of ancient languages. Inscriptions as written records are usable only in proportion to their intelligibility. Important epigraphic corpora remain virtually undeciphered; *e.g.,* the "Indus script" from Mohenjo-daro and Harappā (3rd millennium BC), the Carian texts from Asia Minor and by Carian mercenaries in Egypt (1st millennium BC), and the pictographic Mayan "hieroglyphs" from Central America (*c.* AD 500–1500). Sometimes the writing system is intelligible, as in the case of Etruscan, but understanding remains deficient because the language is otherwise unknown and bilingual keys are lacking or inadequate. Chances for success are best if there are sufficiently extensive bilingual or multilingual copies, of which at least one language is previously understood. Such presence made possible the decipherment of ancient Egyptian (the Rosetta Stone of 196 BC, with hieroglyphic, demotic, and Greek versions). Once the affinity of an underlying language to known idioms is established (*e.g.,* Old Egyptian to Coptic, Old Persian to Avestan and Sanskrit, Akkadian to Hebrew), interpretation can proceed apace. The recovery of Hittite was not a true decipherment because the script was a relatively common variety of syllabic cuneiform. The interpretation was helped by the nature of the writing on the one hand (including intelligible ideograms, while an alphabet yields no such clues), and by the presence of Akkadian-Hittite bilinguals on the other; the soon-recognized Indo-European affinities of the Hittite language afforded further help. The Hittite hieroglyphs were partly deciphered by painstaking internal analysis based on the correct assumption of an underlying dialect akin to Hittite; a bilingual with Phoenician text brought much welcome confirmation. The decipherment of Linear B was a sheer triumph of methodical cryptology, again based on the correct hunch that the hidden language was Greek.

Undeciphered epigraphic corpora

In sum, the decipherment of ancient scripts and the recovery of lost languages are practically identical with the interpretation of previously unintelligible ancient inscriptions, because texts involved are almost exclusively epigraphic. Even when the language is otherwise preserved (as with classical Greek and Latin), inscriptions yield es-

sential additional data for its history, dialects, and social diversification.

History of epigraphy. *Greek and Latin inscriptions.* Inscriptions have commonly elicited the curiosity of posterity, and such ancient Greek historians as Thucydides and Polybius already made scholarly use of them. Sporadic systematic interest in Greek and Latin inscriptions is attested in later ages; *e.g.,* Cola di Rienzo in the 14th century made a collection, and Cyriacus of Ancona (Ciriaco de' Pizzicolli) in the 15th century was a renowned recorder of ancient written monuments on his mercantile travels to Greece, Anatolia, and Egypt. Cyriacus' material formed the nucleus of various compilations in the succeeding centuries, normally on a geographic basis. A rival typological method of publication was launched by Martin Smetius in Leiden in the late 16th century and was followed in the early 17th by a monumental collection of Janus Gruter and Joseph Justus Scaliger. After copious additions of material during the 18th and early 19th centuries, the *Corpus Inscriptionum Graecarum* was launched by August Böckh in 1815 under the aegis of the Berlin Academy and was completed in four volumes with index (1828–77). The material had by then again outrun the publication, and it was resolved in 1868 to re-edit completely all Attic inscriptions. Ulrich von Wilamowitz-Moellendorff in 1902 took charge of the *Inscriptiones Graecae* (1873–), which continued where the *Corpus Inscriptionum Graecarum* left off and included the *Corpus Inscriptionum Atticaru,* as well as all Greek inscriptions from European Greece (including Magna Graecia in Italy) and Cyprus. Those of Anatolia were left to the *Tituli Asiae Minoris* of the Vienna Academy, which began with the Lycian-language inscriptions from Lycia in 1901 and continued with the Greek and Latin ones from Lycia in 1920–44. The rest of Greek Anatolia was combed somewhat by the multivolume American series, *Monumenta Asiae Minoris Antiqua* (since 1928). *Inscriptiones Graecae,* framed in 14 volumes, has turned partly into a kind of overall umbrella for diverse coverage; volumes 6, 8, 10, much of 11, parts of 12, and 13 were never completed, being preempted by such other large publications as *Inschriften von Olympia, Fouilles de Delphes,* V. Latyshev's *Inscriptiones Antiquae Orae Septentrionalis Ponti Euxini Graecae et Latinae* and G. Mikhaylov's *Inscriptiones Graecae in Bulgaria Repertae, Inscriptions de Délos,* and M. Guarducci's *Inscriptiones Creticae.* The continuous influx of new material is inventoried and presented in such annual outlets as *Supplementum Epigraphicum Graecum* (since 1923) and "Bulletin épigraphique" in the *Revue des études grecques.* Large recent undertakings either supplement or complement the *Inscriptiones Graecae,* such as *The Athenian Tribute Lists* (from the American excavations in the Agora) and *Inscriptions grecques et latines de la Syrie* (six volumes, 1929–67). There are, in addition, numerous more localized or specialized collections.

pigraphic ollections

The *Corpus Inscriptionum Latinarum,* founded by Theodore Mommsen in Berlin in 1862, comprises 16 volumes fortified with copious supplements; it covers the classical Latin world and is notable for its homogeneity and systematization. Current finds and studies are presented in *L'année épigraphique* (since 1888). Etruscan inscriptions are less perfectly published in the antiquated *Corpus Inscriptionum Etruscarum;* many important later finds are included in transcription in M. Pallottino's *Testimonia Linguae Etruscae* and in M. Fowler and R.G. Wolfe, *Materials for the Study of the Etruscan Language* (1955; a computerized corpus). Of further relevance to the Roman world are collections of the type of E. Diehl, *Inscriptiones Latinae Christianae Veteres;* M.J. Vermaseren, *Corpus Inscriptionum et Monumentorum Religionis Mithriacae;* and *Inscriptiones Italiae.*

Other inscriptions. R.A.S. Macalister's *Corpus Inscriptionum Insularum Celticarum* (1945–49) gathers the oghamic and other early texts from Ireland and elsewhere. The runic inscriptions are inventoried in a variety of compilations. The *Corpus Inscriptionum Semiticarum,* in Paris (since 1881), covers in separate volumes Phoenician, Aramaic, and other speech areas. Urartean texts were collected by C.F. Lehmann-Haupt in *Corpus Inscriptionum Chaldicarum* (1928–35); the earlier found Hittite hieroglyphic texts, by L. Messerschmidt in the antiquated *Corpus Inscriptionum Hettiticarum.* The *Corpus Inscriptionum Iranicarum,* intended to gather the epigraphs of Persia proper (Achaemenid, Seleucid, Parthian, Sāsānid) and of eastern Iran and Central Asia, began in London in 1955. The *Corpus Inscriptionum Indicarum* has published four volumes since the 1870s, comprising the Aśoka, Indo-Scythian, Gupta, and Kalacuri-Cedi periods, supplemented by the series *Epigraphia Indica* and *South Indian Inscriptions.*

No equally closed corpora exist for cuneiform and Egyptian documents. The Hittite texts come closest to such organization, issued mainly in two great ongoing series, *Keilschrifttexte aus Boghazköi* (German Oriental Society) and *Keilschrifturkunden aus Boghazköi* (Berlin Academy). But mostly the publication is haphazard—as part of excavation records (*Keilschrifttexte aus Assur*), depending on storage sites (*Cuneiform Texts* in the British Museum), or on the basis of genres (F. Thureau-Dangin, *Rituels accadiens*). The same pattern holds true of the Egyptian records.

Much epigraphic material is published in excavation reports and in many archeological and antiquarian periodicals; *e.g., Hesperia, Journal of Hellenic Studies, Bulletin de correspondance hellénique, Archeologia classica, Studi Etruschi,* and *Syria.* Specifically epigraphic serial publications include *Epigraphica* (Milan, since 1939), *Kadmos* (Berlin, since 1962), *Zeitschrift für Papyrologie und Epigraphik* (Bonn, since 1967), and *Chiron* (Munich, since 1971). (J.Pl.)

GENEALOGY

The word genealogy comes from two Greek words—one meaning "race" or "family" and the other "theory" or "science." Thus is derived "to trace ancestry," the science of studying family history. The term pedigree, used to describe a genealogy as set forth in chart or other written form, comes from the Latin *pes* ("foot") and *grus* ("crane") and is derived from a sign resembling a crane's foot, used to indicate lines of descent in early west European genealogies. Chart pedigrees, familiar to most people from school history books, include arrow shapes, parallel lines, a crinkled line denoting illegitimacy, and the sign = denoting marriage. Genealogy is a universal phenomenon and, in forms varying from the rudimentary to the comparatively complex, is found in all nations and periods. In this section the history of genealogy will be outlined, followed by an account of the work of modern genealogists, professional and amateur, and as organized in associations.

History of genealogical study. The history of genealogy can be divided most easily into three stages. The first is that of oral tradition; the second, that in which certain pedigrees were committed to writing. The third stage comprises the period from approximately 1500 in western Europe and later in the English-speaking world, during which the whole basis of genealogy widened to such an extent that it is now possible for the majority of people in western Europe to trace their ancestry.

Oral tradition and biblical sources. In the early days of civilization, before written records were made, oral traditions were necessarily important. Without the art of writing, reliance must be placed on memory, aided possibly by mnemonic systems like that of knot arrangements used by the pre-Hispanic Peruvians, or beads employed by the Maori of New Zealand. The ancient Scottish sennachy, or royal bard, could recite the pedigree of the old Scots kings at the latter's inauguration, and the nobles of Peru, who boasted a common descent with the sovereign, were able to preserve their pedigrees despite the complexity resulting from the practice of polygamy. Oral transmission of genealogical information is almost always as a list of names—the lineages of the ancient Irish kings, for example. Events of outstanding importance are occasionally incorporated in such lists.

Numerous Oriental genealogies appear in the Old Testament. A cursory examination of these will reveal that they belong to the first and second stages in the history

Oriental genealogies

of genealogy, as described above. The systematic keeping of genealogical records, as in Europe since 1500, did not occur until very recently in Asia and Africa.

In southern India the ruling house of the maharajas of Travancore claimed to trace its descent, direct and unbroken, from the old Cēra kings of southern India (referred to as independent sovereigns in one of the edicts of Aśoka, the great Mauryan emperor of the 3rd century BC). A claim that inscriptions of the rulers of Travancore have been found from the 9th century AD comes from a statement issued by the secretariat of the maharaja of Travancore. Its reliability may be judged along with the genealogies of princes in northern India shown in Lt. Col. James Tod's monumental work, *Annals and Antiquities of Rajasthan* (1829, republished 1950). Referring to the lineages of Indian princes as being known since the early centuries BC, Tod wrote, "If, after all, these are fabricated genealogies of the ancient families of India, the fabrication is of ancient date, and they are all they know themselves upon the subject." The very long Oriental genealogies begin as oral pedigrees and were later written down, but they concern only princes or great persons.

In Africa the one instance of a claim to very long descent, that of the Emperor of Ethiopia, bears a similarity to Tod's Rājput genealogies. The Emperor is said to descend from the marriage of King Solomon with the Queen of Sheba. The tradition was written down more than 15 centuries ago; it is therefore older than the history of most European monarchies, but it cannot, of course, be substantiated by documentary proof.

Under European influence the greater Oriental countries have adopted the practice of keeping systematic records for all citizens. In China, with its ancient system of ancestor worship, long, drawn-out pedigrees, including claims to descent from Confucius, are not unknown. The establishment of the Chinese Republic in 1911 brought with it registration of vital statistics.

In modern Japan, as might be expected from its thoroughgoing Westernization, the registration of vital statistics is regulated by law. The Family Registration Law of 1947, and later enactments, require a comprehensive registration of a Japanese national from his birth to his death. Such information, however, is kept in local registration offices, and there is no system in Japan for gathering together, recording, and preserving the information in one central place (although of course the results of statistics, such as the number of births, is known to the central authority). Such an exact system of registration covers only the era of modern Japan. The present-day pedigree of the Japanese emperors has a divine origin; it is mainly a string of names, easily recited and memorized, mixed with semifabulous legends and first written down in the early centuries of the Christian Era. It is concerned only with exalted persons, royal or noble.

Biblical genealogies

In the Old Testament there are many genealogies, the object of which is to show descent from Adam, Noah, and Abraham. By the time these genealogies had become part of the Jewish scriptures, the concept of racial purity had reinforced the keeping of family records. Genealogies of Jesus Christ in the New Testament aim at showing his descent from David, the one in St. Luke's Gospel going as far back as Adam, "who was the son of God." The idea of divine origin was reflected everywhere in a wildly polytheistic form among the Gentiles. Almost without exception, the heroes whose genealogies were recited by the bards had their paternity ascribed to the gods, or to persons such as Romulus who were regarded as having become divine. Greek fables abound in stories of great men begotten by gods and mortals.

In Roman genealogies heroes were always descended from gods. Julius Caesar, for example, was supposed to have sprung from the line of Aeneas, and thus from that of Venus. Among the Romans, traditions of descent remained vague even when written. Caesar's murderer, Brutus, was popularly supposed to be of the same family as an ancient Brutus, who had expelled the Tarquins, but no pedigree appears to have existed to substantiate the belief.

Among the northern nations that overwhelmed the western Roman Empire, belief in divine sonship was general. For Saxon rulers of the English kingdoms it was necessary to be descended from the god Woden.

Early written records. With the invention of writing, the oral became the written tradition. This occurred in Greece and Rome, where genealogies were recorded in poems and in histories. But genealogy did not at this stage become a science, because when writers dealt with it, they did so either incidentally in their narrative or because they were concerned with the family relationships of their gods.

Maximum length of genealogies

The historian Edward Gibbon's observation that "the proudest families are content to lose in the darkness of the middle ages, the tree of their pedigree" may be challenged in the light of recorded genealogies. The male line of Charlemagne has been traced to St. Arnulf, bishop of Metz, who died about 635. Several royal line descents are traceable to the 6th century, as, in England, is the tree of Louis Mountbatten, 1st Earl Mountbatten. The ancestry of Queen Elizabeth II goes to Egbert of Wessex (about 825), beyond him to Cerdic (*c.* 500), and, if another series of names is accepted, to Woden (an actual man later divinized by the Germans), in the 3rd century AD.

With the conversion of the barbarians to Christianity the recording of their regal traditions began. Examples occur in Ireland, Wales, and England. It was natural for the first chroniclers, who were mostly monks, to write down the oral pedigrees of the kings in whose realms they lived. Students of the Irish regal pedigrees are prepared to accept two or three generations before the time of St. Patrick (flourished 5th century AD) as genuine, and it is quite probable that name lists of the Irish kings are valid back to the 3rd century AD. Similarly, in Wales, the ancestry of the greatest Welsh families can be traced for a millennium. Among the Anglo-Saxons there were similar bardic pedigrees recorded by monastic scribes, and many of these might have survived but for the destruction of the Old English ruling class during the Norman Conquest. A regular feature of such old pedigrees recorded by monks was an attempt to link them with the genealogies of the Scriptures. In an Anglo-Saxon pedigree of great length—that of the kings of Wessex (the ancestors of Elizabeth II)—the line is thus traced to Sceaf, "a son of Noah born in the Ark." In the process of working out the connection between scriptural and regal genealogies, the monks adopted a reverse technique to that of the 4th-century-BC Greek mythographer Euhemerus; *i.e.,* they downgraded the old gods to human status.

From roughly 1100 to 1500, the emphasis of genealogists was on pedigrees of royal and noble lines. Claims to a throne, as with the dozen or so claimants to the Scottish crown after the death of Alexander III in 1286 and of his direct heir, Margaret the Maid of Norway in 1290, frequently involved genealogical trees. The truth was sometimes bent to suit some political end, but, on the whole, medieval European records are genealogically valid. This is because they were not primarily intended to supply genealogical information but to record land transactions, taxation, and lawsuits. The facts of family history are incidental and are therefore generally reliable. Exact dates of birth, marriage, and death are rarely given. A man is said to be of age "by Michaelmas 1330."

This period also saw the emergence of pedigrees of lesser folk. Land transactions involved claims in the local courts of the lords. Serfdom gave way to villenage; the latter involved so many days of labour on the lord's demesne and also the inability to move from the estate without the lord's consent. There was strong inducement for a man to prove that he was not a villein and for the bailiff to show that he was. In several parts of England, pedigrees of villeins or persons claimed as such have been worked out over periods of 100–150 years.

It was during the third period in European genealogical history that records that came to include everyone began. This period extends from 1500 to the present. As feudalism gradually gave way, new classes of citizens arose. In England the appearance of a powerful mercantile and business community was reflected in the growth of the middle classes, from which was continually recruited a new nobility and gentry. In turn, owing to the English rule of inheritance by primogeniture and the fact that unlike the

continental nobility English nobility has never extended beyond the reigning peer and his wife, the middle classes themselves continually received the younger children of peers and gentry. Two other factors leading to the proliferation of records were the enormous changes caused by the Reformation and the great reemphasis on individual religion and the desire of Renaissance monarchs to have more exact information about all their subjects.

Modern genealogy. Amateurs in the subject of genealogy are almost always actuated by the desire to trace their own family history. In the course of so doing they discover and work with general principles which apply to pedigrees other than their own, though records other than those applicable to their own case do not interest them. The professional genealogist is concerned not with one family but with many, and with the principles of genealogical research which arise from a wide study. As there are no university courses in the subject and therefore no degrees or other certificates of professional proficiency, the professional must be largely self-taught.

Qualifications of the professional genealogist

The disciplines required of a professional genealogist include a deep knowledge of the history of the country with which he is concerned and of its neighbours. National history determines the form of national genealogy, and genealogy can illuminate many aspects of national history that might otherwise remain obscure. The Wars of the Roses, for example, are hard to grasp unless genealogical trees showing relationships of the contestants are studied, and the course of the American Revolution is easier to understand when the links between George Washington and his compeers with the old English landed families who overthrew the Stuarts are comprehended. An understanding of the principles of law, especially of land law, the ability to decipher court hand or medieval script, an understanding of heraldry, and an intimate knowledge of the study of surnames and place names are also essential to the genealogist. Variations in surname spelling can be bewildering. The key is in the sound of the name, for a medieval scribe could not ask the illiterate person before him to spell his name.

The main task undertaken by professional genealogists is the tracing of pedigrees for clients, this being the staple of their work. Clients often consult genealogists when they wish to establish their family background, or, when having tried to trace it, they have come to a stop.

The writing of private family histories by professionals is very common. The material has usually been worked out by others who wish it to be checked and written by a professional.

Amateur genealogists, as already mentioned, are usually concerned only with their own families. The standard of amateur work varies with the individual, from the truly bad to the excellent.

Amateur genealogical work has increased greatly since 1945. In the United States there has been a long interest in the subject. The New England Historic Genealogical Society, the Augustan Society (based in California) and many state societies are of note. The Mormons (the Church of Jesus Christ of Latter-day Saints) have built up in Salt Lake City, Utah, a microfilm library of genealogical records from Britain and continental Europe, which is probably unequalled. In Canada, Australia, New Zealand, and South Africa the study of genealogy by private persons and by associations is growing rapidly. In England there is a Society of Genealogists, and there are corresponding bodies for Ireland and Scotland. In Denmark there is an International Confederation of Genealogy and Heraldry, which since 1954 has organized international congresses held in many European capitals at intervals of two years. In Czechoslovakia, by way of contrast, the national Genealogical Society has been dissolved, and in general it has not been feasible to obtain genealogical details from Communist countries, though it is probable that changes are now occurring in this respect. Jewish records are in a separate class. With the establishment of the state of Israel in 1948, a very great effort has been made to centralize information about the Jews of continental Europe under the care and direction of The Central Archives for the Study of the Jewish Peoples, in Jerusalem.

In tracing family history, the worker follows certain rules. He works backward from the present. This is an elementary caution constantly put on one side by amateurs, who tend to trace forward from a person of the same name who may well be unrelated. As there cannot in the nature of things be a gap in a pedigree, no assumptions as to relationship can be allowed without very strong reason to accept them. Good and bad features in the ancestry have to be accepted: bastardy has to be allowed for, as well as regal ancestry by legitimate lines. An ancestor's wrongdoing must also be allowed for, though with passage of time this is usually taken in a romantic sense. Registration of birth, marriage, and death first became compulsory in England in 1837. Public records in most other Western countries began at varying dates in the 19th century. Census records are of great importance. They began in the United States as early as 1791; in Britain in 1801 (papers kept only from 1841); and even earlier in French Canada, in 1655–66. Parish registers began in England in 1538, though they are rarely preserved from that date. In most countries they begin later, but in Spain the oldest extant is dated 1394, and there are 1,636 parishes having records prior to 1570. In England, Nonconformist records have been kept by various bodies, and many are now held officially at Somerset House, London, or the Public Record Office. In America the settlers were generally trying to get away from established church and controlling state. They were vigorous individualists who kept careful records of their lives and of the organization of their new communities. Examples of detail in New England records can be seen in many of the 1,600 pedigrees contained in *American Families with British Ancestry* (500 page suppl. to Burke's *Landed Gentry,* 1939), in A.M. Burke's *Prominent Families of the United States of America,* and in the many volumes of family history produced in the United States.

Main genealogical sources

Wills are of the utmost importance as a source of genealogical information. Ships' lists of passengers are useful in supplying dates for immigrant ancestors' departure for the New World, but since they do not indicate place of origin, but only the port of departure, the original habitat must be sought elsewhere. Without knowledge of the ancestor's place of origin in the homeland it is useless in almost all cases to attempt a search.

With the aid of the type of record mentioned above, and with help from family Bibles, tombstones, and plaques and brasses in churches, it is as a rule possible for a person of English antecedents to trace some 250–300 years of ancestry. Before the 17th century, everything depends on the social position of the ancestors. Tax records, lawsuits, and purchases and sales of land are the chief sources for tracing a family before 1600. The Pipe Rolls extend from the reign of Henry II (1154–89) to that of Queen Victoria (1837–1901), with an interrupted beginning also in the time of Henry I (1100–35). Monastic records are of great importance as showing grants or ownership of land. The pleas of the crown deal with suits at law and contain much detail about families. There are many Rolls besides those of the Pipe that give a great deal of incidental genealogical information. *Inquisitiones post mortem* show the position of an heir; *i.e.,* his age and other details. As the centuries are passed, the numbers of those who can prove a descent by the male line dwindle, until by the time of the Norman Conquest scarcely half a dozen pedigrees can be traced in the male line for either Saxon or Norman.

Regarding deposition of public records, two principles have been followed by archivists: that of centralization and that of diffused local holdings. The former has many obvious advantages and was adopted in Scotland and in Ireland. It has one disadvantage—destruction of the records at one stroke. This happened in Dublin on April 13, 1922, when Irish factions fighting with each other burned most of the Irish records. The second system, by which records are stored in a number of depositories, prevails to a considerable extent in England. Although the Public Record Office, Somerset House, and the British Museum library are places of centralized record, the parish registers remain outside them, scattered in numerous parishes or county offices. County records contain masses of material not to be found in London.

From the 16th century there has been an increasing accumulation of written material, which deals either exclusively or incidentally with genealogy. William Camden (1551–1623), a learned English antiquary and historian, did much to raise the standards of genealogical research. He was the first English writer on surnames, and his work was not resumed for nearly 200 years. Sir William Dugdale, a younger contemporary of Camden, made a beginning with his *Antiquities of Warwickshire* to the great output of county histories written between the 17th and 19th centuries. The revolution in modern genealogy was the application to its study of canons of historical and literary criticism formulated in Europe from 1800. Their application to genealogy was fairly late, as is illustrated by the fact that the 19th-century English historian Thomas Macaulay, critically perceptive in most other spheres, accepted what amounted to family myths as true genealogy. Later writers, including J.H. Round, W. Farrer, C. Lewis Loyd, C.T. Clay, and the editors of the *Complete Peerage*, are of the greatest importance. (L.G.P.)

PALEOGRAPHY

Paleography (from Greek *palaios,* "old," and *graphein,* "to write") is the study of ancient and medieval handwriting. Precise boundaries for the study are hard to define. For example, epigraphy, the study of inscriptions cut on immovable objects for permanent public inspection, is related to paleography. Casual graffiti, sale or election notices as found on the walls of Pompeii, and Christian inscriptions in the Roman catacombs are likewise part of paleographical knowledge. In general, however, paleography embraces writing found principally on papyrus, parchment (vellum), and paper. Today, paleography is regarded as relating to Greek and Latin scripts with their derivatives, thus, as a rule, excluding Egyptian, Hebrew, and Middle and Far Eastern scripts. It is closely linked with diplomatic (see above), the study of forms in which official and private documents are drawn up.

The scientific study of Latin paleography (and diplomatic) dates from 1681, when the French monk Jean Mabillon published *De Re Diplomatica,* the first textbook on the subject, while his compatriot Bernard de Montfaucon performed a parallel service for Greek paleography in his *Palaeographia Graeca* in 1708.

The primary task of the paleographer is to read the writings of the past correctly and to assign a date and place of origin. Close acquaintance with the language of the text is a prerequisite. Help in dating is offered by changes in styles of handwriting and variations from area to area. Abbreviations in texts likewise help in dating and localization.

Types of writing materials. A paleographer must be familiar with writing materials. Any smooth surface able to accept writing has served in the past, notably, pottery fragments, animals' shoulder blades, slabs of wood, bark, cloth, and metal.

Papyrus

The great writing material of the ancient world was papyrus, in use by 3500 BC. In preparing the surface, strips taken from the papyrus reed (*byblos*) growing in the Nile Delta were laid side by side, while other strips were laid across at right angles, and the whole impregnated with paste. After treatment a fine smooth surface was obtained. Much of the administration of the Roman Empire depended upon papyrus, in the same way that modern bureaucracies depend upon paper. Warfare and a damp climate resulted in an almost total disappearance of papyrus from Europe, though the dry (if war-scarred) sands of Egypt have preserved vast numbers of documents. Papyrus was imported into Europe from Egypt even after the fall of Rome. Chance survivals include charters of Merovingian kings in France (7th century) and business documents (5th–10th centuries) at Ravenna, the old administrative capital of the late Roman Empire.

The other great ancient writing material, still in occasional use today, is parchment, or vellum, the terms being often used interchangeably. Vellum is a term usually applied to skin from a calf (*cf.* veal, *veau*), while parchment is an expression often applied to sheepskin or goatskin. The word parchment is derived from Pergamum in Asia Minor, the ancient centre of its manufacture.

Both papyrus and parchment were expensive and were replaced for everyday use by wax tablets corresponding to today's notebook. Tablets made of wooden blocks were hollowed out and filled with melted, often black wax. Notes were made in the hardened surface. Even documents of permanent significance, such as property conveyances, were made on wax tablets.

Because ancient writing materials were expensive, they were often reused. Papyrus presented difficulties, for ink soon bonded itself firmly into the surface. Parchment could be more readily reused, because it is tougher and can be washed or scraped clean. Many medieval monks, when short of writing materials, took ancient books to pieces, cleaned off the leaves and used them again. The original script can often be brought out under ultraviolet light. Parchments thus cleaned and freshly inscribed are called palimpsests (Greek *palin,* "again"; *psēstos,* "scraped").

Paper is the third great writing material. In use in China at a remote period, it was employed extensively in the Arab world by the 9th century. Not in common use in Europe until the 14th century, it took over the name of the half-forgotten papyrus.

In the early classical world the standard form of book was the papyrus roll, commonly called *biblion,* taking its name from the material of which it was made. It consisted of papyrus sheets pasted edge to edge with a slight overlap. The text was set out in columns, drawn up at right angles to the edge of the rolls, and started at the left. The reader unrolled as he went along and at the conclusion was obliged to reroll the book. The roll was an inconvenient form of book, difficult to consult, which probably accounts for the inaccurate quotations found in early literature, caused by an author relying on his memory rather than troubling to unwind a long roll. By the time of Christ, a new form of book was coming into fashion, the codex, or book in the shape in which it is known today. The codex is almost always of parchment, since papyrus cracks when folded. The codex seems first to have been used for notebooks or account books, the conservatism of booksellers and readers ensuring the survival of the roll for centuries. The Christians popularized the codex, using it for the Gospels.

Form and binding of texts

Various instruments have been used for writing. The early Egyptians used a slender rush. From about 300 BC the thicker reed pen was used. The reed was in general use in the Greco-Roman world. Metal pens, copied from the reed, were also employed. For wax tablets a stylus was used, made of wood, bone, ivory, iron, or bronze. In many northern European areas, where reeds suitable for writing purposes are not indigenous, the feather (*penna*) became the main writing instrument. It was usually stripped of its vanes and the quill alone used.

Writing implements

Lead, used in classical times for ruling guidelines in manuscripts, was used extensively in the Middle Ages for rough notes and annotations in the margins of books.

Ink has been prepared in a variety of ways. In classical times the black discharge of cuttlefish was used, as well as concoctions of soot and gum. In the Middle Ages oak apples were steeped in water with "vitriol" (ferrous sulphate) to produce ink.

Analysis of texts. The essential skill of a paleographer is the ability to recognize the numerous styles of handwriting prevalent in different ages and places. Most European scripts descend from Greek and Roman capital letters, but variations are enormous. It is a European convention that writing starts on the left at the top and works line by line down the page. An eccentricity known as boustrophedon (from Greek *boustrophēdon,* "following the ox furrow"), whereby alternate lines are written backward in mirror writing, occurs chiefly in very ancient inscriptions.

The Greek and Latin alphabets existed originally as capital, or majuscule, letters. The ancient Greek alphabet, as developed in chiselled inscriptions on stone or marble, served without much modification as the alphabet used in literary works written on papyrus rolls. This script, found in the oldest surviving Greek literary papyri of *c.* 300 BC or earlier, gave way to more rounded and elegant

forms, probably developed in the Greek literary circles of Alexandria. Cursive scripts that were easier to write were developed for everyday use, for business, and to record the acts of the great bureaucracy of Egypt, where the Greeks settled in large numbers. The Greek cursive script and the formal book script greatly influenced each other, as can be seen from a vast series of cursive documents dating from the 4th century BC for about 1,000 years. Because so much material survived, early Greek cursive can be better studied than its Latin counterpart. In Greek cursive manuscripts the everyday life of ordinary people becomes a reality: they pay or fail to pay taxes, buy or sell houses, and harass civil servants with awkward demands.

A very rough division in Greek paleography may be made at around AD 300. The earlier age is called the papyrus period; and the later, the parchment or Byzantine (or Christian) period. The division, however, is imprecise, for parchment was used well before and papyrus long after this date. The change from papyrus to parchment is signalized by three great monuments of paleographical studies, the Vatican, Alexandrine, and Sinai Bibles, all on parchment and in codex form.

An alphabet of small, or minuscule, letters developed gradually and was in use by the 8th century. Numerous abbreviations exist in Greek manuscripts, though never so many as in Latin. Accents, an additional complexity, were not systematically applied before the 7th century AD.

The ancient Latin alphabet of capitals (*quadrata*) is found in numberless inscriptions in stone and marble all over the Roman world. How far this alphabet was used for writing books is uncertain, because, though excellently adapted for incision, it is difficult to write. Some specimens of handwriting in *quadrata* do exist, such as 4th- or 5th-century copies of Virgil, but scholarly opinion largely regards these as abnormal productions. By the 1st century a handsome Latin alphabet existed, called rustic, based on the use of a broad pen or brush. Rustic was used for public inscriptions on walls, as in the sale and election notices found at Pompeii. Although specimens are scarce, it is likely that books were extensively written in this hand in classical times. By the 4th century another Latin alphabet existed, the script known as uncial, in the nature of a rounded form of *quadrata.* Uncial survived the fall of Rome and from it developed half-uncial, the ancestor of the small letters in use today.

Styles of writing

The stately Roman scripts, *quadrata,* rustic, or uncial, were not used for everyday purposes, and, as in the case of Greek, a cursive, rapidly written hand arose in which letters and business documents were inscribed. This hand is found in graffiti on Pompeian walls and in wax tablets. After the disintegration of the empire, Roman cursive became the ancestor of regional hands in what are now Spain, France, and Italy.

During the flowering of Christianity and art in Ireland (*c.* 500–*c.* 1000) a beautiful "insular" script developed, which found its way into England. There, two streams of influence commingled, for from 597 Christian missionaries arrived from Rome and brought in books in uncial script. Both scripts prospered in England, though insular gradually superseded uncial.

The most successful of all scripts proved to be Caroline minuscule, which takes its name from the emperor Charlemagne (died 814), patron of scholars and scribes, under whom the script was developed. Despite its inherent superiority and clarity, it did not predominate over regional scripts until the mid-12th century, and the local hand of southern Italy (Beneventan) maintained itself for much longer.

In the 12th century, Caroline minuscule, which had undergone moderate developments, started to display more obvious changes. It compressed laterally, while its rounded strokes became stiffer and straighter as it was converted into the so-called Gothic hands—very angular in northern Europe and more rounded in Italy. A revulsion against Gothic took place in scholarly circles in Italy in the 14th and 15th centuries, and a return to models based on Caroline minuscule took place. This revived hand, called Humanistic because humanist scholars used it, was adopted by 15th-century Italian printers, whose type faces ultimately triumphed over the Gothic. (This encyclopaedia is printed in a type scarcely modified from Caroline minuscule.) Meanwhile, Caroline and Gothic scripts had produced cursive hands for quick everyday use, as in the case of the ancient Greek and Latin alphabets. These cursive scripts were used for the vast mass of business documents written in the Middle Ages.

Cursive scripts

Abbreviations are the principal problem confronting paleographers. They were extensively used in Roman times by lawyers to avoid repetition of technical terms and formulas. Abbreviations fall into two classes, suspension and contraction. Suspension, omission of the end of a word and indication by a point or sign, was used in Roman public inscriptions—*e.g.,* IMP.(ERATOR), CAES.(AR). Contraction, the omission of letters from the middle of a word and replacement by a sign or some other device, was common among Greek-speaking Jews, who contracted certain sacred or revered names, such as God, Lord, Israel, or David, as a mark of veneration. The Christians followed the practice by contracting their sacred names, such as Jesus and Christos. The great increase in use of abbreviation as a means of saving time and material dates from the 12th century, but some contraction signs are of high antiquity, such as the sign 7 for *et* (Latin, "and"). Some are quickly written versions of letter groups, such as "÷" for *est* (Latin, "is"), the top dot standing for e, the bottom dot for t, and the stroke being a long or f-shaped s, fallen on its side. The letter r is often omitted, the adjacent vowel being written above the line, as c^{a}ta for *carta* ("charter"). In works for semilearned readers, such as romances, abbreviations are often few, but books produced for the learned, such as university textbooks, are heavily loaded with abbreviations. The number of signs and devices in use by the end of the Middle Ages was enormous. More than 13,000 are listed in the standard work, Adriano Cappelli's *Lexicon Abbreviaturarum* (1912).

Abbreviations

Dating of books and documents also offers problems. Even when a precise date is given, the dating system of a given time and a given area must be checked because the year began at different times in different territories, and there were even variations in the same country. Calendar reforms initiated in 1582 by Pope Gregory XIII, for example, were not adopted in Protestant England until 1752. If a year of a monarch's reign is given as a date, it is necessary to determine whether the reign is counted from his accession or coronation. Moreover, few books were dated, the dated title page being nonexistent in medieval works, though sometimes a final paragraph, the colophon, supplies a date with the scribe's name and place of work. Important documents, such as English 12th-century royal charters, are undated.

In the absence of dates, inferences are drawn from handwriting, use of abbreviations, and internal evidence. Caution must be used, however; for an elderly scribe may be using a hand learned over half a century before, and work in the same scriptorium or office as a young clerk anxious to show off all the latest tricks and flourishes. Some styles lasted a long time: Caroline minuscule lasted for more than three centuries. Certain kinds of books, such as liturgical volumes, were produced in a highly stylized form for generations, and thus it is often difficult to provide a close date for a late medieval missal (with standard illustrations and marginal decorations) in a mechanical Gothic hand.

Internal evidence must be weighed carefully. A given historical event noted in a chronicle will provide an earliest possible date, unless the entry is an interpolation. Evidence for some legal practice or liturgical usage is no safe guide, for legislation on the subject by a king or a pope may merely be ratification of a long-standing practice.

A paleographer must get to know his scribes, for their mannerisms can be highly informative. Nearly 50 different scribes have been distinguished in the English royal chancery in the period 1100–89. First-rate scribes, such as notaries public, provide much information about themselves, giving their names, notarial signs, and information on their authority to act. Even anonymous clerks, who drew up innumerable property conveyances, can be identified by their script over a period of years, and their career

Internal evidence

can be traced through developing, mature, and deteriorating handwriting, thereby offering dating evidence.

The provenance (origin) of many manuscripts can be immediately recognized because certain centres developed individual styles. Papal and royal chanceries issued documents of easily identifiable origin, while many monastic scriptoria—for example, that of Canterbury Cathedral in the earlier 12th century—had a virtually private handwriting.

Textual corruptions are another obstacle to correct elucidation. A legal document is certain to have been checked at the time of writing, but one cannot be sure in the case of a literary, philosophical, or theological text. Scribes were fallible, and, if there are no signs of any corrections in a text, then it probably embodies inaccuracies. A popular book, such as Chaucer's works, exists in large numbers of manuscripts, and many manuscripts produce variant readings. If a scribe made a mistake in copying, future scribes using his version are likely to reproduce the error and add others. Sometimes the same muddled passage in a group of manuscripts of a given author can be traced back to damage in an earlier copy, say a section eaten by rodents or impenetrably stained. Whenever copyists worked from different and faulty originals, various copies tend to fall into families. A paleographer must bring together various readings in families and decide which is the best reading.

Sometimes a scribe, set to work because he could write a fine hand, did not necessarily possess much knowledge of the language. Such a scribe faced with a text heavily loaded with abbreviations would usually make nonsense of it. Occasionally, a particularly stupid copyist, faced with a master copy in two columns of writing, would copy straight across the top line, then across the second. When he used a different number of words per line, the text was reduced to unintelligibility. In Greek and Roman times there was the difficulty that texts were written continuously, without space between the words. Copyists misread passages. For instance the historian Tacitus reported that some tribesmen went off to guard their own property: ADSVATVTANDA (*ad sua tutanda*). Some copyist thought "Suatutanda" was a place and this ghost name was perpetuated in geographical works. Later medieval Gothic hands presented a forest of vertical strokes called minims. The letter *v* rendered as *u* made two strokes, while *i* was often left without a dot or at best with a faint hairline, often misplaced. The group of letters *ium* could be read, as *uim, uiui, niui, mui, miu,* with many other variations. Accordingly, minim corruption, confusion of vertical strokes, is a term constantly heard in paleographical circles.

Latin and Greek are inflected languages in which the same case and tense endings constantly occur, offering scope for error. Moreover, in biblical, theological, or philosophical texts, the same words abound. For example, in one place in the Gospel According to John there occurs the passage:

> Verba quae ego loquor
> vobis a me ipso non loquor
> pater autem in me manens . . .

("The words that I speak . . ."). The eye of a sleepy scribe might slip from the first *loquor* to the second, whereupon he would go on copying at *pater autem,* leaving out the second line altogether, a common type of error known as *homoioteleuton* ("like ending").

Because of the lack of surviving specimens, it is difficult to assess book decoration in classical times, but apparently it was very limited. In the later centuries of the Roman Empire, however, book illustrations were not infrequent.

Decorations

The narrative material in the Bible encouraged illustration. The Irish were foremost in applying decoration to the text in the form of elaboration of capital letters, producing such masterpieces as the Book of Kells (late 7th century), in which Celtic imagination and artistic sense ran riot in elevating the book to an object of outstanding beauty. Some of the greatest creative talent of the Middle Ages was lavished upon books, especially upon those used in worship, such as Bibles, psalters, and missals. When a book cannot be assigned either a date or provenance upon the appearance of the text alone, its style of illumination will often direct the paleographer to a certain monastery in which the carving on capitals or wall paintings may contain the same motifs.

Because of the immensely high prices of manuscripts, the question of forgery naturally arises, but it is safe to say that no modern forgery could survive for a moment. A convincing imitation of ancient script is virtually impossible, while the papyrus, parchment, or paper on which it would be written could not stand up to modern scientific inspection. Anything of recent vegetable or animal origin fluoresces brightly under ultraviolet light, to name but one test. William Henry Ireland (died 1835), the Shakespeare forger, used flyleaves from 16th-century books, but his handwriting and non-Shakespearean language gave him away. A modern would-be forger must either copy an existing work, which, in the present state of art history and paleographical study, would be immediately recognized, or be prepared to invent medieval subject matter.

Forgeries

There was fabrication of documents in medieval times on a considerable scale. A monastery might find itself in possession of estates held since remote antiquity but without any title deeds. When some powerful monarch made difficulties, there was a strong inducement to produce the required ancient-looking documents. The borderline between justifying legitimate possession and culpable attempts to gain extra territory or privileges, however, is ill-defined. Monks occasionally descended to falsifications of title deeds and charters of exemption. About 1125 a monk of Soissons on his deathbed confessed to a career of professional forgery for gain and admitted fabricating charters for various monasteries, including Westminster Abbey. Early forgeries, however, give themselves away through such inconsistencies as mentioning bishops of nonexistent sees or embodying legal phrases that came into use generations later or bearing seals when seals were not yet appended to documents.

The modern paleographer has great technical aids: photography since the 19th century and colour photography in the 20th. Ultraviolet light brings out faded handwriting. Microfilm makes the contents of a volume in a far-distant repository available quickly and cheaply. (W.G.U.)

SIGILLOGRAPHY

Sigillography is the study of seals. A sealing is the impression made by the impact of a hard engraved surface on a softer material, such as clay or wax, once used to authenticate documents in the manner of a signature today; the word seal (Latin *sigillum;* old French *scel*) refers either to the matrix (or die) or to the impression. Seals are usually round or a pointed oval in shape or occasionally triangular, square, diamond, or shield-shaped.

Medieval matrices were usually made of latten—a kind of bronze—or of silver. Ivory and lead were occasionally used, gold very rarely. Steel was used from the 17th century. Matrices could include intaglio gems. The usual material for the impression was sealing wax, made of beeswax and resin, often coloured red or green. In southern Europe, notably in the papal Curia, lead and occasionally gold were used. Shellac, the wax used today, was introduced in the 16th century.

Purposes of seals

Seals were used to establish the authenticity of such documents as charters and legal agreements and for the verification of administrative warrants. In southern Europe, early medieval documents were drawn up by notaries and authenticated with their written *signa,* but this never replaced seals in northern Europe. Forgeries were manufactured as early as the 12th century, indicating how important seals had become. From that time, also, seals were used to close folded documents and thus to guarantee their secrecy. Seals were also used to affirm assent; for example, by a jury. Under the Statute of Cambridge (1388), sealed letters were used in England for the identification of people and their places of origin.

Sigillography is used to assist other historical studies. Many impressions have survived from the medieval period. Those attached to documents are most valuable, because the documents may date their use precisely and the seal may confirm the documents' authenticity. Unattached seals may still provide useful evidence from their inscription or design. Fragmentary seal impressions are often

difficult to interpret. Fewer matrices have survived and often without related impressions. Seals often reflect the taste of the owner. They provide evidence for changes in fashion in both secular and ecclesiastical costume and for the development of armour. Seals indicate heraldry before the earliest rolls of arms and are an original source for armorial bearings, which thereby enables the historian to trace the distinctions or alliances between various families and so contributes to genealogy. Craftsmen's seals often display tools connected with trade. Depictions of towns, churches, castles, and monasteries, although conventional, can often aid the architectural historian. Seals can also be used profitably for studying ancient ships, particularly their shape and details of masts and rigging. The main difficulty in studying seal designs is that they were often conservative, especially since seals were often replaced with an exactly similar design.

(J.Ch.)

Seals in antiquity. Seals with designs carved in intaglio were used throughout antiquity. They were of two main types—the cylinder and the stamp. The cylinder first appeared in Mesopotamia in the late 4th millennium BC and continued to be used there until the 4th century BC. It was also widespread in Elam, Syria, and Egypt (3rd millennium BC) and in Cyprus and the Aegean (2nd millennium BC). Stamp seals preceded cylinders, first appearing in Mesopotamia in the 5th millennium BC and developing over a period of about 1,500 years until largely replaced by the cylinder in the 3rd millennium. Early stamp seals were also used in Iran, northern Syria, and southeastern Anatolia during the 4th and 3rd millennia; in Anatolia their use was continued in the 2nd millennium by the Hittites. In Mesopotamia the stamp seal gradually came into use again in the 8th–6th centuries, effectively replacing the cylinder by the 3rd century BC. In Egypt the scarab largely replaced the cylinder seal early in the 2nd millennium BC and continued as the main type until replaced by the signet ring in Roman times. In the Aegean, various types of stamp seals were used throughout the 2nd and much of the 1st millennium BC, until in Hellenistic and Roman times the signet ring became dominant.

The uses of ancient seals are known from textual references and ancient sealings, both on lumps of clay and on documents found in excavations. In historical times most prominent citizens, including women, carried their own seals. That the rank or office of the owner was often included in the inscription indicates that many of these may have been official seals. Kings had their own seals, and high officials could hold the king's seal as a mark of delegated authority.

Uses of ancient seals

Seals came into use before the invention of writing for the securing of jars, bales, bags, baskets, boxes, doors, etc., and this use continued throughout ancient times. The method was either to shape clay over the stopper or lid or to make a fastening with cord and place clay around the knot and then impress it with the seal.

The sealing of written documents, of which the two major ancient classes were clay tablets and papyrus scrolls, became regularly established in the latter part of the 3rd millennium BC. The clay tablet was the main vehicle of writing in Mesopotamia, where cuneiform was used into the Christian Era, and this method spread to Elam, Anatolia, Syria, and Egypt; clay tablets were also used in the Aegean Bronze Age. The tablet was inscribed while the clay was soft, and seal impressions were applied at the same time.

The main kinds of sealed cuneiform documents were contracts, accounts, and letters. Contracts were sealed by the contracting parties and by witnesses and commonly encased in clay "envelopes," on which the text was either repeated or summarized and the seals again impressed. Two special kinds of contract were the royal grant to a subject, impressed with the royal seal of the grantor, and the treaty between nations, a number of examples of which have been recovered, some of them bearing impressions of the seals of the royal contracting parties. Account tablets were sealed to authenticate the transfer of goods. A letter was often encased in an "envelope" and the sender's seal impressed on the outside to identify him to the recipient and, in the case of official letters, to authenticate any commands contained in the letter.

In Egypt, papyrus documents may be assumed from soon after 3000 BC, but surviving evidence dates mostly from the latter part of the 3rd millennium onward. The method of sealing a papyrus document was to roll it into a tube, tie a strand or cord around the centre, and seal a clay lump over the knot. This method continued into the Christian Era, from which time a great number of Greek papyri have survived in Egypt. It was not until the 1st millennium BC that this kind of document, including by then leather and parchment, came into wide use outside Egypt. The spread of this kind of document, on which the space for a seal was small, probably played some part in the gradual replacement of the cylinder by the stamp seal.

No documents or sealings have been discovered from ancient India, but the still undeciphered inscriptions on the seals may include personal names, perhaps of merchants, who could have used the seals in much the same ways as their Near Eastern contemporaries, with whom they are known to have had commercial contacts.

Since seals were used throughout ancient times and are sufficiently durable to have survived in very large numbers, they form one of the few classes of ancient objects in which a continuous development can be traced. The great majority bear artistic representations, so their chief value is for art history, but, since these include details of environment (plants, animals), equipment (plows, chariots, musical instruments), or dress, they also contribute to cultural history.

Further information is provided by the inscriptions on seals. The existence of rulers known only from king lists may sometimes be confirmed by the discovery of their seals, and in some cases rulers are known only from their seals, which, because they often mention the names of their fathers, the cities that they ruled, and the chief gods that they served, form a valuable historical source. The assembling of tablets and sealings bearing the impressions of private seals can contribute to the reconstruction of business archives and the destinations of traded goods, thus providing valuable material for economic analysis, and far-flung trade contacts can be deduced from foreign seals in excavations (*e.g.*, Indus Valley seals in Babylonia). Personal names are an important source for ethnic analysis, and inscribed seals, because they often name the owner's father and even grandfather, provide material for this as well as for genealogical reconstruction. (T.C.M.)

Medieval European seals. The connection between Roman and medieval seals lies in the use of seals in the chanceries of the Merovingian and Carolingian kings. Many Ottonian seals had busts of the emperors. Royal seals of medieval type, with the ruler enthroned and bearing his insignia, appear from the 11th century. The use of seals by bishops and nobles became usual at this time and was widespread by the 12th century. By the 13th century, seals were used by all classes, including small landowners; and, by the 14th century, simple seal matrices could be bought ready-made.

Spread of sealing in medieval times

The quality of engraving varied greatly. Some delicately designed seals date from the 12th century, such as the silver seal of Isabella of Hainaut, queen of France in 1180–90. The silver equestrian seal of Robert Fitzwalter is a notable example of the 13th century, the period of the finest seal engraving.

The names of several engravers of medieval seals are known: for example, Luke, who engraved the seal of Exeter, and Walter de Ripa, who engraved the first great seal of Henry III of England.

Forms of medieval seals. Seal matrices may be single or double, thus producing an impression on either one or both sides of the wax. Single matrices, the older type, often have a ridge along the back and end in a loop. Double matrices, known from the 11th century onward, are flat, with two to four projecting lugs pierced with holes in which vertical pins keep the halves aligned.

Sealing both sides of the wax makes detaching the seal more difficult, and so in medieval times the reverse was often sealed by a counterseal for greater security. The official seal of an institution was often countersealed by

the seal of an official, such as a town by its mayor. Single seals were often fitted with a handle; the most common type was a six-sided cone ending in a trefoil. In some matrices the centre screwed outward, enabling the device to be used without a legend. On many seals the back was marked with a cross to indicate the top.

Seals could be either applied to the surface of a document or appended from it by a strip of material. Application was the earlier system, although papal bulls were always appended. Appended seals appeared in England in the 11th century and in France in the 12th; seals were appended either on a tongue of parchment cut across from the bottom of the document or on a tag of parchment, leather, or silk inserted through a cut in the document. Some documents had many seals. Seals were often protected by woven bags or by boxes of wood, metal, or ivory known as skippets.

The legend

The legend, often abbreviated, usually declared the name of the owner or institution; it often began with a cross and the word *sigillum,* followed by the name in the genitive case. Latin remained in fashion for inscriptions, though English and French are occasionally found from the 13th century, more frequently on personal seals. On English seals roman capitals were used in the 11th and 12th centuries and Lombardic ones in the 13th and early 14th centuries. Black letter (Gothic script) was first used in England in the 14th century and was quite popular in the 15th century, although Lombardic often continued for capitals. Roman capitals reappeared in the 16th century.

Royal and official seals. The great seal, or seal of majesty (a round seal showing the seated ruler with the royal insignia), first appeared in Europe on the seal of the emperor Henry II of Germany (ruled 1002–24), in France on the seal of Henry I (ruled 1031–60), and in England on the double seal of Edward the Confessor (ruled 1042–66). The seal of William I of England (ruled 1066–87) had the King on one side and an equestrian figure on the other. The kings of France adopted double seals under Louis VII (ruled 1137–80).

The development of lesser royal seals can be illustrated by the growth of English government. Deputed great seals were used for the major legal courts and for France, Ireland, and Wales. The expansion of the kings' affairs caused the addition of smaller, more personal seals, such as the signet. The Chancery did not control these seals, and this freedom led to the evolution of autonomous offices. The privy seal appeared early in the 13th century in the custody of the clerks of the king's chamber. It was soon transferred to the wardrobe clerks, and gradually its importance increased until by the early 14th century the keeper of the privy seal was the third minister of state. The keepership gained further prestige in midcentury, when the great seal was entrusted to the keepers who went abroad with Edward III. As the privy seal grew in importance, the king preferred another small seal for authenticating correspondence and warrants. Under Edward II (ruled 1307–27) there was a secret seal distinct from the privy seal. By 1400 the signet, as the secret seal was then called, was in the charge of the king's secretary. The signet rather than the privy seal became the originating force in administration, and from 1540 there were two secretaries, each with two signets. The privy seal and signet seal were both single armorial seals.

Royal officials had their own seals. Circular admirals' seals, dating from the late 14th century to the 17th century, include a fine group of 15th-century bronze matrices. The seals show ships in great detail, with the sails displaying the arms of the admiral.

Religious seals. The principal episcopal seal was the seal of dignity, always a pointed oval. From the 11th to the 14th century it usually depicted the standing figure of the bishop, from the 13th century with a canopy above him. In the mid-14th century the standing figure was often replaced by a saint or a religious scene, with the bishop praying beneath—a form that had been used earlier on episcopal counterseals. The seal of Thomas Arundel, archbishop of Canterbury (1396), depicts the martyrdom of Becket in the centre of an elaborate series of niches, with the archbishop below.

Bishops' and monastic seals

Monastic seals, usually double-sided and of high quality, normally show the buildings of the monastery, religious scenes, or the patron saint. They were distinct from abbots' and priors' seals, which were similar to those of bishops. Notable was the elaborate four-part matrix of Boxgrove Priory (mid-13th century). The seal of Merton Priory (1241), considered the finest English medieval seal, had the Virgin and child on one side with St. Augustine of Hippo on the other.

Papal bulls were doubled-sided lead seals appended to the document on strings. The earliest known is that of Deusdedit (reigned 615–618). The usual design, with the head of the Apostles Peter and Paul on one side and the pope's name on the other, first appeared under Paschal II (reigned 1099–1118). Although this style of portrayal of the heads was changed in the Renaissance, the design has not been altered.

Town seals. The possession of a common seal was an important part of a town's independence. Town seals were almost always round and often double. Many towns still possess their original matrices. The earliest in England date from around 1200, when many towns received their charters. The seal of Exeter has been dated to *c.* 1180. Maritime towns often depicted a ship with a furled sail; inland towns often showed the guildhall or the town itself. The seal of Rochester depicted the Norman castle within a wall. Counterseals often bore the figures of saints. Later medieval town seals were less common and beginning with the 14th century were often single seals.

Commercial seals. Commercial seals in England were considerably increased during the reign of Edward I, when double-sided bronze dies occurred for various customs, subsidies, and the delivery of wool and hides. Their obverse displayed the arms of England, and the reverse had the same device without the shield. The seals of merchants and craftsmen often displayed either merchants' marks or tools connected with their trade.

Personal seals. The earliest class of personal seals was that of greater barons of the 12th century, who used the device of a fully armed equstrian knight. Their shields often provide the earliest evidence for the use of heraldic charges. Some greater barons used a double seal with an equestrian obverse and an armorial reverse. The most usual type of personal seal was a single seal with the arms of the owner. Women's seals were usually pointed ovals and showed the lady standing, sometimes between shields. Nonheraldic personal seals displayed a variety of devices, such as stars, fleurs-de-lis, armorials, and religious subjects. The inscription sometimes indicated the owner, although it may simply have related to the device, as on those that bore the device of a squirrel and the inscription "I krack nuts."

Modern use of seals. The use of seals declined as the use of signatures grew. Personal fob seals were in fashion from the 17th to the early 19th century; often they were gems in gold settings that were carried in the fob or breeches pocket and were used to seal folded private correspondence before the envelope was introduced. States and institutions continued to use seals for the formal ratification of their acts, but few of these seals maintained the medieval vigour of design. (J.Ch.)

Chinese and Japanese seals. The private seals used in China (*t'u-chang*) and Japan (*ingyō*), commonly square and reading merely "seal of so and so" (XX *chih yin*), served as a confirmation of signature or a sign to be verified but have not the legal status of a signature. They are made of ivory, wood, or jade. Used by artists and collectors to mark their paintings and books, there is hardly a limit to their fanciful designs and phraseology. A man might own scores of seals, using his many sobriquets, especially those suggesting unworldly and rustic tastes. A seal is impressed in red ink—made of cinnabar in water and honey or suspended in sesame oil, hempseed oil, etc.—held ready on a pad of cotton or moss. The characters most often appear in line, but they are sometimes reserved against the inked ground.

The first record of a seal in China is from 544 BC. Actual bronze seals survive from the 5th century BC, and the practice of sealing must be some centuries older. The emblematic characters cast on Shang dynasty bronze vessels

(13th–11th century BC) imply the use of something like a seal for impressing on the mold. The royal seal and other seals of high office were termed *hsi;* other seals of rank and appointment were *chang.* The imperial *hsi* (called *pao* beginning in the T'ang period, AD 618–907) was traditionally large and square, made of jade or ivory. The most famous one belonged to Shih Huang-ti (ruled 221–209/210 BC); it had as its knob a one-horned dragon and is fabled to have been handed down to the present day.

"Small" and "great" seals

The official and, no doubt, the personal use of seals began in Japan with the copying of Chinese institutions in the 7th century AD. Both in China and in Japan modern seals generally employ the "small seal" character (*chuan shu*), the "great seal" character being reserved in the past for the ruler and high officers. To the historian the importance of the Far Eastern seals is greater in the earlier periods, and in China they yield more information than in Japan. Thus, seals recovered archaeologically throw light on government appointments made in the Han period, particularly in the reign of Han Wu Ti (140–87 BC), when they were tokens of rank given to internal officials and some external client rulers. Gold seals of the "King of the Han Wei-nu county," found near Fukuoka in 1784, and that of the "King of Tien" excavated near K'unming in 1956 have implications of this kind. But in post-Han times the seals have served little if at all as primary historical documents, and in writings on East Asia it is chiefly the art historian who appeals to their testimony in authenticating paintings and calligraphies.

(W.W.)

TEXTUAL CRITICISM

Textual criticism defined

The technique of restoring texts as nearly as possible to their original form is called textual criticism. Texts in this connection are defined as writings other than formal documents, inscribed or printed on paper, parchment, papyrus, or similar materials. The study of formal documents such as deeds and charters belongs to the science known as "diplomatic"; the study of writings on stone is part of epigraphy; while inscriptions on coins and seals are the province of numismatics and sigillography. Textual criticism, properly speaking, is an ancillary academic discipline designed to lay the foundations for the so-called higher criticism, which deals with questions of authenticity and attribution, of interpretation, and of literary and historical evaluation. This distinction between the lower and the higher branches of criticism was first made explicitly by the German biblical scholar J.G. Eichhorn; the first use of the term "textual criticism" in English dates from the middle of the 19th century. In practice the operations of textual and "higher" criticism cannot be rigidly differentiated: at the very outset of his work a critic, faced with variant forms of a text, inevitably employs stylistic and other criteria belonging to the "higher" branch. The methods of textual criticism, insofar as they are not codified common sense, are the methods of historical inquiry. Texts have been transmitted in an almost limitless variety of ways, and the criteria employed by the textual critic—technical, philological, literary, or aesthetic—are valid only if applied in awareness of the particular set of historical circumstances governing each case.

The value of textual criticism

An acquaintance with the history of texts and the principles of textual criticism is indispensable for the student of history, literature, or philosophy. Written texts supply the main foundation for these disciplines, and some knowledge of the processes of their transmission is necessary for understanding and control of the scholar's basic materials. For the advanced student the criticism and editing of texts offers an unrivalled philological training and a uniquely instructive avenue to the history of scholarship; it is broadly true that all advances in philology have been made in connection with the problems of editing texts. To say this is to recognize that the equipment needed by the critic for his task includes a mastery of the whole field of study within which his text lies; for the editing of Homer (to take an extreme case), a period of some 3,000 years. For the general reader the benefits of textual criticism are less apparent but are nevertheless real. Most men are apt to take texts on trust, even to prefer a familiar version, however debased or unauthentic, to the true one. The reader who resists all change is exemplified by Erasmus' story of the priest who preferred his nonsensical *mumpsimus* to the correct *sumpsimus.* Such people are saved from themselves by the activities of the textual critic.

The law of diminishing returns operates in the textual field as in others: improvements in the texts of the great writers cannot be made indefinitely. Yet a surprisingly large number of texts have not yet been edited satisfactorily. This is particularly true of medieval literature, but also of many modern novels. Indeed the basic materials of most textual investigation, the manuscripts themselves, have as yet not all been identified and catalogued, much less systematically exploited. The first edition of the works of Dickens to be founded on critical study of the textual evidence did not begin to appear until 1966, when K. Tillotson's edition of *Oliver Twist* was published. Reliable principles of Shakespearean editing have begun to emerge only with modern developments in the techniques of analytical bibliography. The Revised Standard Version of the Bible (1952) and the New English Bible (1970) both incorporate readings of the Old Testament unknown before 1947, the year in which early biblical manuscripts—the so-called Dead Sea Scrolls—were discovered in the caves of Qumrān.

The materials of the investigation. The premise of the textual critic's work is that whenever a text is transmitted, variation occurs. This is because human beings are careless, fallible, and occasionally perverse. Variation can occur in several ways: through mechanical damage or accidental omission; through misunderstanding due to changes in fashions of writing; through ignorance of language or subject matter; through inattention or stupidity; and through deliberate efforts at correction. The task of the textual critic is to detect and, so far as possible, undo these effects. His concern is with the reconstruction of what no longer exists. A text is not a concrete artifact, like a pot or a statue, but an abstract concept or idea. The original text of Aeschylus' *Agamemnon* or Horace's *Odes* has perished; what survives is a number of derived forms or states of the text, approximations of varying reliability preserved by tradition. The critic must reduce these approximations as nearly as possible to the first or original state that they imperfectly represent; or if, as sometimes happens for reasons that will be explained below, no single original can be reconstructed or postulated, he must reduce their number to the lowest possible figure. His methods and the degree of his success will be determined by the nature of the individual problem; *i.e.,* the text itself and the circumstances of its transmission. The range of possible situations is vast, as the following survey indicates. The types of text with which the critic is concerned may be classified broadly under three heads.

Books transmitted in print. For practical purposes it is often assumed that the latest edition of a modern book published during the author's lifetime may be treated as the original. This is a simplification. The actual author's original may have been a manuscript or a typescript or a recording; in the process of publication it has passed through several stages of transmission, including possibly storage in a computer, at any one of which errors have necessarily occurred. Experience teaches that some errors will survive uncorrected in the published version. Further errors are likely to occur if a book is reprinted. Even an edition revised by the author is not to be regarded as textually definitive. Errors committed and overlooked by the author himself may be corrected by the critic in appropriate cases. Special problems are posed by an author's second thoughts, whether preserved in his books and papers or incorporated in editions revised by him; recent research has shown that authorial revision in modern printed books has been underestimated. The extent to which a critic is free to choose between authorial variants on aesthetic grounds is a matter of debate.

Books published before the 19th century pose essentially similar problems in a more intractable form, as may be seen in the case of Shakespeare. No manuscript of any of Shakespeare's plays survives, and there were substantial intervals between the dates of composition and the first

printed versions, in which unauthorized variation clearly occurred. For Shakespeare's plays, indeed, the very concept of an author's original may be misleading. Elizabethan printers clearly had little regard for strict textual accuracy, so that allowance must be made not only for error but for deliberate alteration by compositors; thus the textual criticism of 16th- and 17th-century books must include a study of the practices of early printers.

Books transmitted in manuscript. Nearly all classical and patristic texts, and a great many medieval texts, fall into this category. Every handwritten copy of a book is textually unique, and to that extent represents a separate edition of the text. Whereas the characteristic grouping of printed texts is "monogenous" (*i.e.,* in a straight line of descent), that of manuscript texts is "polygenous" or branched and interlocking. The critic is in principle obliged to establish the relationship of every surviving manuscript copy of a text to every other. The difficulty and indeed the feasibility of this undertaking varies enormously from case to case. The following extremes embrace a wide range of intermediate possibilities. (1) The authority for a text may be a single surviving copy (*e.g.,* Menander, *Dyscolus*) or a copy that can be shown to be the source of all other copies (*e.g.,* Varro, *De Lingua Latina*) or an edition printed directly from a copy now lost (*e.g.,* the work of the Roman historian Velleius Paterculus); or a text may be transmitted in scores of copies whose interrelationships cannot be exactly determined (*e.g.,* Claudian, *De raptu Proserpinae*). (2) The interval between the original and the earliest surviving copies may be very short (*e.g.,* the French medieval poet Chrétien de Troyes) or very long (*e.g.,* the Attic tragedians). (3) A tradition may be "dynamic"—*i.e.,* the text may have been copied and recopied many times even in a short time (*e.g.,* Dante's *La divina commedia*); or it may be "static"—*i.e.,* the number of transmissional stages even over a long period may have been few (e.g., *Epigrammata Bobiensia,* a Latin translation of Greek epigrams). (4) A text may be a religious or literary work that was respectfully treated by copyists and protected by an exegetical tradition (*e.g.,* the Bible, the Latin poet Virgil); or a popular book that was exposed to correction, glossing, and amplification by readers (*e.g.,* the *Regula magistri* ["Rule of the Master," a Latin work related to the Rule of St. Benedict] and much medieval vernacular literature). (5) A text may have been written and transmitted after the establishment of a scholarly tradition, or it may show signs of "wild" and arbitrary variation dating from an age in which standards of exact verbal accuracy were low. To this extent all Greek books written before the establishment of the Alexandrian library (see below) were exposed to the hazards associated with oral transmission.

Books transmitted orally. Many texts have been orally transmitted, sometimes for long periods, before being committed to writing, and much textual variation may be attributable to this stage of transmission. Often in such cases the critic cannot attempt to construct an "original" but must stop short at some intermediate stage: thus the edited text of Homer means in practice the closest possible approximation to the text as established by the scholars of Alexandria. The length, complexity, and fidelity of oral traditions varies enormously. The text of the old Indian Rigveda was transmitted orally almost without variation from very ancient to modern times, whereas much old French epic and Provençal lyric has descended in variant redactions for which a common source may be postulated but cannot be reconstructed. Sometimes this is attributable not to spontaneous variation but to deliberate reworking, whether by the author, as appears to be the case with the three versions of the English poem *Piers Plowman,* or by later revisers, as with the four versions of *Digenis Akritas* (a Greek epic). The distinction, however, is not always easy to draw. These considerations apply to a wide range of texts from ancient Hebrew through Old Norse to modern Russian, but they are especially important for medieval literature. In this field perhaps more than in any other the critic's aims and methods will be dictated by the character of the oral tradition, the stage at which it attained a more or less fixed form in writing, and the attitude of copyists in a particular genre to precise verbal accuracy. A problem of particular difficulty and importance is posed by the Greek New Testament. Though the text appears to have been transmitted from the first in writing, the textual variations are in many ways analogous to those of an oral tradition, and it is commonly held that the essential task of the critic is not to try to reconstruct the "original" but to isolate those forms of the text that were current in particular centres in the ancient world.

Critical methods. From the preceding discussion it is apparent that there is only one universally valid principle of textual criticism, the formulation of which can be traced back at least as far as the 18th-century German historian A.L. von Schlözer: that each case is special. The critic must begin by defining the problem presented by his particular material and the consequent limitations of his inquiry. Everything that is said below about "method" must be understood in the light of this general proviso. The celebrated dictum of the 18th-century English classical scholar Richard Bentley that "reason and the facts outweigh a hundred manuscripts" (*ratio et res ipsa centum codicibus potiores sunt*) is not a repudiation of science but a reminder that the critic is by definition one who discriminates (the word itself derives from the Greek word for "judge"), and that no amount of learning or mastery of method will compensate for a lack of common sense. To study the great critics in action is incomparably more instructive than to read theoretical manuals. As the editor of Manilius, A.E. Housman, wrote,

The importance of common sense

> A man who possesses common sense and the use of reason must not expect to learn from treatises or lectures on textual criticism anything that he could not, with leisure and industry, find out for himself. What the lectures and treatises can do for him is to save him time and trouble by presenting to him immediately considerations which would in any case occur to him sooner or later.

Admittedly, the technical advances in textual bibliography mentioned below are not such as would sooner or later occur to any reflective and intelligent person; but bibliography, like paleography, is ancillary to textual criticism proper, and Housman's words are strictly true. What they imply is that good critics are born, not made.

The critical process can be resolved into three stages: (1) recension, (2) examination, and (3) emendation. Though these stages are logically distinct, (2) and (3) are in practice performed simultaneously, and even (1) entails the application of criteria theoretically appropriate to (2) and (3).

The search for the earliest transmitted form of the text

Recension. The operation of recension is the reconstructing of the earliest form or forms of the text that can be inferred from the surviving evidence. Such evidence may be internal or external. Internal evidence consists of all extant copies or editions of the text, together with versions in other languages, citations in other authors, and other sources not belonging to the main textual tradition. These witnesses (as they may be called) must be identified, dated, and described, using the appropriate paleographical and bibliographical techniques. They must then be collated; *i.e.,* the variant readings that they contain must be registered by comparison with some selected form of the text, often a standard printed edition. Where the number of witnesses is large, collation may have to be of selected passages. If there is only one witness to a text, collation and recension are synonymous, and the critic passes straight to examination and emendation. Generally, however, he will be faced with two or more witnesses offering variant forms or states of the text.

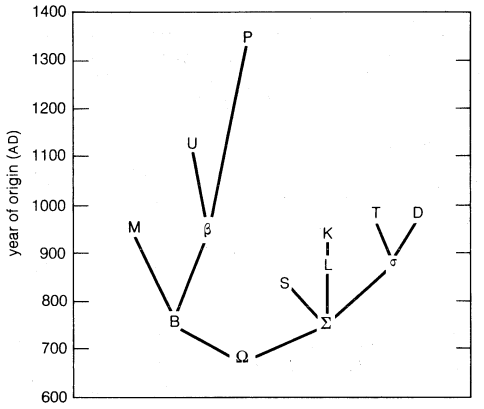
Collateral evidence as to the transmission of a text may be supplied from sources external to the direct or indirect textual tradition. Thus the ancient biographers throw light on the circumstances in which Virgil's *Aeneid* was published. Inferred textual stages may be dated on the evidence of copying practices at different periods, or by association with a particular scholar, or from entries in medieval library catalogues. Generally speaking, information of this sort will contribute more to the history than to the criticism of the text, but the two fields are intimately connected; and the better the textual history is known, the more reliable the control of the critic over conjectural solutions to specific problems. In the case of

printed books, such external evidence is as a rule more plentiful; it is often essential, since so much may turn on the accurate dating of editions. Relevant information must be sought in the published and unpublished records of stationers, printers, booksellers, and publishers and in other archival material.

Having assembled his evidence, the critic may proceed, broadly speaking, in one of two different ways, according as he decides to handle the problem of interrelationships "genealogically" or "textually."

In the "genealogical" or "stemmatic" approach, the attempt to reconstruct an original text here relies on the witnesses themselves regarded as physical objects related to each other chronologically and genealogically; the text and the textual vehicle (the book itself) are treated as a single entity. On the basis of shared variants, chiefly errors and omissions, a family tree of the witnesses (*stemma codicum*) is drawn up. Those witnesses that repeat the testimony of other surviving witnesses are discarded, and from the agreements of the remainder the text is reconstructed as it existed in the lost copy from which they descend, the "archetype." Thus in the tradition of the 6th-century monk Cassiodorus' *Institutiones* the relationships of the manuscripts of the authentic version of the text of Book II may be represented by the accompanying diagram. The Roman letters represent extant manuscripts, and the Greek letters represent the lost manuscripts from

Adapted from R.A.B. Mynors (ed.), *Cassiodori Senatoris Institutiones* (1961); the Clarendon Press, Oxford



Manuscript genealogy, with Greek letters representing sources upon which known manuscripts (Roman letters) must have been based.

which they derive, here arbitrarily dated. The text of the archetype Ω is established by the agreement of B and Σ. Since B survives, the readings of MUP, which are derived from it, would be of value only where B had suffered damage after M and β were copied from it. In such cases the text of β could be inferred from the agreement of UP and the text of B from the agreement of Mβ (or MU or MP). The text of Σ can be inferred from the agreement of SLσ or SL or Sσ (or ST or SD) or Lσ (or LT or LD). K, being copied from L, would be of value only where L had suffered damage after K was copied from it. An important distinction is here exemplified between "trifid" and "bifid" *stemmata.* Where there are three independent witnesses to a source, as with Σ, its reading is certified by the agreement of all three or of any two; where there are only two witnesses, as with Ω, and they disagree, the reading of the source cannot be certified. Even in the latter situation, however, the number of possible variants existing in the source would have been reduced to two. Thus in theory the genealogical, or stemmatic, method allows the critic to eliminate from consideration all variants that cannot be traced back to the archetype or earliest inferable textual state.

While in principle this method is unassailable, it depends for its practical validity on the assumption that each copyist followed only one model or exemplar and generated only variants peculiar to himself. This is called "vertical" transmission, and a tradition of this kind is called "closed." Once the possibility is admitted that a copyist used more than one exemplar or (the more probable supposition) copied an exemplar in which variants from another source or sources had been incorporated—*i.e.,* that more than one textual state may coexist in a single witness—the construction of a *stemma* becomes more complicated and may be impossible. This is called "horizontal" transmission, and a tradition of this kind is called "open" or "contaminated." The practice of critics faced with contamination tends to vary, for historical reasons, from field to field. Editors of classical texts generally adopt a controlled eclecticism, classifying the witnesses broadly by groups according to the general character of their texts and choosing between their readings largely on grounds of intrinsic excellence. Medievalists, following the French scholar Joseph Bédier (see below), sometimes revert to the traditional practice, to which their training may dispose, of selecting a single witness as the main basis of the text. For editors of printed books, contamination is not an important problem.

Analysis of variants

In the "textual" or "distributional" approach, the text and the textual vehicle are dissociated; the emphasis is on the analysis of the variants themselves and their distribution rather than on the character of the text as presented by individual witnesses. The techniques or models employed include those of statistics, symbolic logic, and biological taxonomy. Two theoretical advantages are suggested for this approach. First, objectivity: no judgments of value are entailed, whereas the genealogical method calls for decisions as to the correctness of readings or textual states. Second, the possibility of mechanization: long and elaborate calculations involving thousands of variants may be performed by a computer. This possibility is especially attractive to New Testament critics, who are confronted with about 5,000 manuscripts of the Greek text as well as versions in other languages and patristic citations. In practice, however, these advantages are to a large extent illusory. An "objective" (*i.e.,* undiscriminating) treatment of all variants in a literary text such as Ovid's *Metamorphoses* (of which more than 300 manuscripts exist) without regard to their metrical and stylistic quality would be a self-evident waste of time and produce merely confusion. The critic cannot abrogate his critical function, which implies discrimination, at the very beginning of the critical process. Moreover, the preparation or programming of a text for treatment in this way, whether mechanical aids are used or not, is long and laborious, and one must consider whether in a given case the results justify the expenditure of effort. Texts have been transmitted by a combination of purpose and accident that in any particular instance is both unique and unpredictable, and no machine or statistical model exhibits the versatility necessary to unravel the incomplete and tangled skein. Mechanical methods have been most successful in fields other than recension (see below).

Authenticity of a text

Examination. The process of determining whether the transmitted text or any of the transmitted variants of it is "authentic"—*i.e.,* what the author intended—is known as examination. The prior process of recension has reduced the number of textual states having a claim to be considered "authoritative." Many different situations are possible. In a completely closed tradition it is theoretically feasible to reconstruct the archetype with such certainty that only a single form of the text without variants remains to be examined. In practice this is extremely unlikely to be the situation. Usually the critic is faced with pairs (sometimes triplets) of variants, all with a presumptive claim to be considered authoritative. In some traditions he will confront variant versions of the whole text. Where papyri or other early sources independent of the main tradition are available, he may have to reckon with "pretraditional" (*i.e.,* pre-archetypal) variants. The process of examination calls upon the critic's full range of knowledge as well as his innate powers of taste and discrimination. The criteria applied must be those appropriate to the particular author (supposing his identity to be known), the period, the genre, and the particular character of the work. The opposing demands of analogy and anomaly must be weighed according to the circumstances. Many of the older generation of critics based their decisions on aprioristic or rigidly

analogical principles of elegance and propriety, while the canons of modern criticism are based on historical studies of language and style. It is here that the circularity inherent in the whole operation is most evident, for the linguistic and stylistic criteria employed are themselves based on inductions from texts, probably including the one under examination. There is no escape from this difficulty; as the German philologist Karl Lachmann observed, it is precisely the task of the critic "to tread that circle deftly and warily."

Emendation. The attempt to restore the transmitted text to its authentic state is called emendation. There will usually be a chronological gap, sometimes of several centuries, between the archetype, or earliest inferable state of the text, and the original; nearly all manuscripts of classical authors date from the Middle Ages. The history of the text during the intervening period may be illustrated from external sources; but if examination has convinced the critic that the transmitted text (or its variants) are not authentic, he normally has no recourse but to bridge the gap by conjecture. Conjectural emendation has been defined by the American scholar B.L. Gildersleeve as "the appeal from manuscripts we have to a manuscript that has been lost." Theoretically this definition is acceptable, if we interpret "manuscript" as "source," but in practice the making of conjectures, as distinct from testing them, is intelligent guesswork.

Conjectural emendation

No part of the theory of textual criticism has suffered more from misunderstanding than has conjectural emendation. Such conjectural, or divinatory, criticism has in the past enjoyed a traditional preeminence: Dr. Johnson observed that William Warburton's correction of "good" to "god" in the second act of *Hamlet* (scene 2, line 182) almost set the critic on a level with the author. That idea is as erroneous as the frame of mind in which the Italian scholar C. Pascal founded the Paravia series of editions in order to purge Latin texts of German conjectures. The best critic is he who discriminates best, whether between variants or between transmitted text and conjecture.

Conjectures as a rule occur to the mind spontaneously or not at all; diagnosis and prescription often present themselves at the same moment. This instinctive process is not under the critic's control, though he can sharpen and regulate it by constant study and observation. The outcome of the process, the emendation itself, can and must be controlled and tested by precisely the same criteria as are used in deciding between variants. This is essentially an exercise in balancing probabilities. These probabilities are historical. The conventional distinction between intrinsic and transcriptional (*i.e.,* paleographical or bibliographical) probability tends to obscure a fundamental historical point. If the transmitted form of the text lies at few removes or a short distance in time from the original, a conjectural solution which violates transcriptional probability is less likely to be correct than if the text has undergone a long and complex process of deterioration. In the latter case the critic may attach little or no importance to transcriptional probability. The critic cannot neglect the study of paleography or bibliography, but he must not give them more than their critical due. What that may be depends on the particular historical circumstances. He will study carefully the rationale of error in manuscripts and books themselves rather than in the schematic classifications of critical manuals; and he will learn from experience to distinguish between the types of error that may be called "psychological" (*i.e.,* those committed by a tired or inattentive copyist, whatever language or instruments he uses) and those contingent on the period and the medium of transmission, whether it be the mouth and the ear, the pen, the hand composing stick, the linotype or typewriter keyboard, the computer or photocopying machine, or the printing press. Two complementary principles originated by the New Testament critics of the 18th century are often cited as aids to decision: *utrum in alterum abiturum erat?* ("which reading would be more likely to have given rise to the other?") and *difficilior lectio potior* ("the more difficult reading is to be preferred"). These are no more than useful rules of thumb; it has been suggested that in practice these and other such principles reduce themselves to the truism *melior lectio potior,* "the better reading is to be preferred."

From this discussion it is apparent that the traditional opposition between "conservative" and "radical" styles of criticism that has haunted textual criticism since St. Jerome has no meaning. The critic does not attack or defend the transmitted text; he asks himself whether it is authentic. How radically he treats it, and how many conjectural readings he substitutes for transmitted readings, depends not on his temperament but on the nature of the problem. If he has studied the history of textual criticism he will know that as a matter of demonstrable fact nearly all conjectures are wrong, and he will accept that many of his solutions are in the nature of things provisional.

Editorial technique. Critical texts are edited according to conventions that vary with the type of text (classical, medieval, modern) but follow certain general principles. In some cases, as with newly edited papyri and with palimpsests (writing materials re-used after erasure), the edition will take the form of a diplomatic transcript; *i.e.,* the most accurate possible representation of a particular textual form. Generally, however, the editor constitutes his text in accordance with his own judgment on principles explained in his introduction; and he indicates his sources in critical notes (*apparatus criticus*), preferably at the foot of the page. These notes are usually couched in a special terminology that relies heavily on abbreviation and the use of conventional signs or letters (*sigla*) to identify the witnesses. In classical and patristic texts the language of the notes is usually Latin. Editorial judgment will be influenced by the presumed needs of readers: in an edition intended for scholars, very corrupt passages are often printed as transmitted and marked with a dagger (†), whereas in an edition for the student or general reader some compromise may be accepted in the interests of readability.

A much-discussed problem is the treatment of "accidentals"—variations in spelling, capitalization, punctuation, and the like. Few if any ancient text traditions preserve reliable evidence of authorial practice in these matters, so that the editor is concerned only with variants that affect the sense; in preparing his text for printing he will adopt modern conventions of presentation and punctuation and a normalized orthography. The same holds good for the majority of medieval texts. Printed texts, however, were generally corrected or seen through the press by the author, or at all events by a contemporary, so that the editor may be reasonably confident of reproducing at least a decent approximation to authorial usage. Whether, or to what extent, he should do so is much debated; opinions differ sharply as to the usefulness of "old-spelling" editions of Shakespeare and other early writers.

History of textual criticism. *From antiquity to the Renaissance.* Until the 20th century the development of textual criticism was inevitably dominated by classical and biblical studies. The systematic study and practice of the subject originated in the 3rd century BC with the Greek scholars of Alexandria. Literary culture had before that time been predominantly oral, though books were in common use by the 5th century, and many texts had suffered damage because the idea of precise textual accuracy and reproduction was unfamiliar. The aim of the librarians of Alexandria was to collect and catalogue every extant Greek book and to produce critical editions of the most important together with textual and interpretative commentaries. Many such editions and commentaries did in fact appear. Alexandrian editing was distinguished above all by respect for the tradition; the text was constituted from the oldest and best copies available, and conjectural emendation was rigidly confined to the commentary, which was contained in a separate volume. An elaborate battery of critical signs was used to refer from text to commentary. These techniques were applied, though on a less ambitious scale, by Roman scholars to Latin texts. Fidelity to tradition was the chief legacy of ancient textual scholarship to later ages; the copyist was expected to reproduce his exemplar as exactly as he could, and correction was based on comparison with other copies, not on the unaided conjectural sagacity of the scribe. Such was the practice

The librarians of Alexandria

of the best monastic scriptoria such as that of Tours, or of the best scholars, such as Lupus of Ferrières (fl. 850). From about 1350, however, a change in attitude is evident, particularly in the West. What is often called the revival of learning was in reality a practical movement to enlist the heritage of classical antiquity in the service of the new Christian humanism. In order to make them usable (*i.e.*, readable), texts were corrected freely and often arbitrarily by scholars, copyists, and readers (the three categories being in fact hardly distinguishable). At its best, as seen in the activities of a scholar like Demetrius Triclinius, later medieval and early Renaissance criticism verges on scientific scholarship, but such cases are exceptional. For the most part the correction of texts was a purely subjective display of taste, sometimes right but much more often wrong, and resting as a rule on nothing more solid than a superficial sense of elegance. In consequence, by the 1470s, when the first printed editions (*editiones principes*) of classical texts began to appear, most Greek and Latin authors were circulating in a textually debased condition, and it was manuscripts of this character that almost always served as copy for the early printers. Very little editing in any real sense of the word was done; the scholars who saw the *editiones principes* through the press generally confined themselves to superficial improvements.

Debasement of texts during the Renaissance

From Politian to Cobet. This state of affairs entailed that down to the 19th century most critics were engaged not in establishing and emending texts on scientific principles but in correcting, in a necessarily unsystematic fashion, a vulgate or received text (*lectio recepta*) that was itself the product of an almost entirely haphazard process of variation and conjecture. The situation was aggravated by the fact that the manuscripts themselves, the basic materials of the investigation, were largely inaccessible to scholars. The Italian poet and scholar Politian, unlike most of his contemporaries, was aware that only through the identification and comparison of the best manuscripts could texts be improved; his notes and collations show that he understood the problem correctly as essentially one of control of the sources. What might have been done in this field is shown by his work, cut short by his early death, on the Florentine codex of Justinian's Pandects. Many manuscripts were still privately owned, their very existence unknown to scholars; public libraries were few and published catalogues fewer; travel was difficult, expensive, and often dangerous. It was not until the twin disciplines of diplomatic and paleography were founded by the great Benedictine monks Mabillon and Montfaucon, and developed by their successors, that a critical use of the evidence became possible; and much of the evidence itself did not become available until after the Napoleonic Wars, when most of the private stock of manuscripts passed finally into public collections.

Some advances were taking place, slowly and unsystematically, in both the theory and practice of textual criticism. The history of critical method in this period is most profitably studied in the best editions of the best editors. The accepted method was to correct the text (*i.e.*, the text of the last printed edition) *codicum et ingenii ope; i.e.*, with the aid of the manuscript and printed sources and the critic's own ingenuity. Divination was subordinated to authority, and any reading found in a manuscript or printed text was accounted superior to any conjecture, whatever its intrinsic merits. The first important departure from this pattern is seen in the edition of Catullus by J.J. Scaliger (1577), in which the possibilities of the genealogical method, already understood in principle by Politian and other Renaissance scholars, were exemplified by the demonstration that all the extant copies derived from a lost manuscript, whose orthography and provenance Scaliger was prepared to reconstruct.

Richard Bentley's influence

Almost equally significant was Richard Bentley's edition of Horace (1711), in which for the first time the role of conjecture in the critical and editorial process was recognized and the tradition of producing a corrected version of the text of previous editors was decisively rejected. Bentley's scholarship was greatly admired in the Netherlands, and the editions of the great Dutch Latinists J.F. Gronovius and N. Heinsius were informed by Bentleian principles. Under his influence there grew up what may be called an Anglo-Dutch school of criticism, the two most typical representatives of which were Richard Porson and C.G. Cobet. Its strength lay in sound judgment and good taste rooted in minute linguistic and metrical study; its weaknesses were an excessive reliance on analogical criteria and an indifference to German science and method. Its influence may still be seen in the empiricism that characterizes much critical work by English scholars.

From Bentley to Lachmann. The decisive influence on the editing of secular texts came from the New Testament critics of the 18th century. The printed text of the Greek New Testament in common use was still essentially that established in 1516 by Desiderius Erasmus. For his edition, produced in great haste, he had used such manuscripts, neither ancient nor good, as chanced to be accessible to him. Superficially revised, this was the text termed in the Elzevier edition of 1633 "now received by all," *nunc ab omnibus receptum.* Bentley proposed an edition on radical lines in which he engaged to give the text "exactly as it was in the best exemplars at the time of the Council of Nice. So that there shall not be twenty words, nor even particles, difference . . ." This project never materialized, but editions of the Greek text that did not reproduce the *textus receptus* were published in England by Daniel Mace (1729), William Bowyer, the Younger (1763), and Edward Harwood (1776). On the Continent, meanwhile, New Testament criticism was being developed on scientific and historical lines by a succession of distinguished scholars, notably J.A. Bengel, J.J. Wettstein, J.S. Semler, and J.J. Griesbach. They shaped the genealogical method that was later refined by editors of classical texts. Wettstein also deserves commemoration as the first New Testament critic to use *sigla* systematically. This was important, since some at least of the deficiencies of classical editions at this time are attributable to the lack of suitable conventions for the presentation of critical information, together with a conservative and belletristic attitude to technical jargon by publishers, scholars, and users of books in general. Though *sigla* occur sporadically in editions as early as the 16th century and were used by S. Haverkamp in his Lucretius (1725) in something like the modern style, they did not become normal until the second half of the 19th century.

The genealogical, or stemmatic, method of recension has already been described. It is usually associated with the name of the German Karl Lachmann, but it had its origins in the work of J.A. Bengel and his successors, and almost every essential feature of it was already present in the work of Lachmann's precursors such as J.A. Ernesti, F.A. Wolf, K.G. Zumpt, F.W. Ritschl, and J.N. Madvig. Nevertheless Lachmann occupies a central position in the development of textual criticism because of the unusual power and penetration of his scholarship, the range of textual material on which he worked, and his immense contemporary and posthumous influence. His edition of the Greek New Testament (1831; 2nd ed. 1842–50) was intended primarily as a vindication of the principles of Bentley and Bengel and a demonstration that the *textus receptus* must be finally rejected. Similarly his famous edition of Lucretius (1850) is important as an exemplification of the method in action, since the tradition of Lucretius is peculiarly suitable for the purpose. The demonstration fell short of completeness, for Lachmann had not fully grasped the problem and so failed to exploit the method fully. It has been suggested that Lachmann's best critical work was in his editions of medieval German texts; their influence will be considered below. The Lachmannian model of recension derived added authority from seemingly analogous models in other fields, especially that of comparative philology. As propagated by disciples, notably Moritz Haupt, it dominated textual studies for half a century.

The genealogical method of Karl Lachmann

Related developments in the late 19th century. Possibly the most important technical advance in the latter part of the 19th century was the perfection of photography. Instead of travelling in search of his material, the paleographer or critic could now assemble and study it at relatively little expense and without leaving his desk.

During the last quarter of the 19th century the tempo of archaeological discovery in classical and biblical lands

was vastly increased, and many new texts were unearthed. Some of these were in previously unknown languages, setting new problems of decipherment. Specifically relevant to textual studies are the many Greek papyri recovered from Egypt. These have thrown much light on the history and techniques of ancient book production and scholarship and hence, indirectly, on critical problems. Where the texts they contain are already known, their evidence has tended to emphasize our ignorance of the textual history of classical literature in antiquity itself. Being usually far older than the manuscripts already known, they often illuminate the "pretraditional" state of the text; by sometimes offering readings that agree with those of late and "inferior" medieval copies they justify editors in a policy of cautious eclecticism. Papyrus discoveries have been of particular moment for the text of the New Testament.

Editors of printed texts, having invariably received a classical education (no other being available), had naturally followed, with minor modifications, the methods of classical editing. They would reprint the text of the last edition with such improvements as editorial taste and learning suggested but with no attempt to investigate the sources of the text. Since Lachmann's method was inapplicable to printed texts, this procedure continued until, by the end of the 19th century, the text of Shakespeare, for example, was in a state somewhat analogous to that of most classical writers at the time of the *editiones principes.* Much of the work of modern Shakespearean editors has consisted of undoing the damage inflicted by their predecessors. The early 20th century saw the rise of a new school of "biblio-textual" criticism, most notably represented by A.W. Pollard, R.B. McKerrow, and W.W. Greg. Its object was to devise a style of recension appropriate to the special circumstances under which early printed texts were produced and propagated, and its methods were those of analytical bibliography. These developments are of direct importance for the criticism and editing of a large range of texts of the 16th, 17th, and 18th centuries, particularly those of the Elizabethan and Jacobean dramatists. They have also engendered a discussion of general methodological interest on the role of bibliographical as opposed to historical and literary criteria in the editorial process. This debate continues.

Shakespeare texts

Critics and editors of medieval texts had also inevitably been influenced by developments in the classical field. Before Lachmann it had been usual to choose a single manuscript as the main basis for an edition. Because of the circumstances in which much medieval literature was composed and transmitted this was not necessarily unscientific, and the surviving bulk of texts was so large as to dictate that approach in many cases if they were to be edited at all. This had been the style of editing followed by the Belgian Jesuits known as Bollandists, the French Benedictines called Maurists, and the Italian scholar L.A. Muratori, and perpetuated in the indispensable *Patrologiae Cursus Completus* (edition of the Church Fathers) of the French priest Jacques-Paul Migne. At its best it is seen in the editions of medieval Latin chronicles by the 18th-century Oxford antiquary Thomas Hearne, some of which are still standard works. A more scientific approach was adopted in the publications of the *Monumenta Germaniae Historica,* the later volumes of which (from about 1880) were produced by editors trained in the school of Lachmann. Similarly, editors of vernacular texts followed the lead that Lachmann had given in his editions of such early German poems as the *Nibelunge Not* (1826) and the *Iwein* (1827). An important development in the application of the method was due to the medievalists G. Gröber and G. Paris, who first emphasized the significance of common errors. But in the general uncritical enthusiasm for scientific method, the genealogical approach was too often used without regard for the special conditions under which medieval literature has been handed down.

Reaction against the genealogical method. Haupt had proclaimed in his lectures that his main object was to teach method. But confidence in method led to its misuse. The Lachmannian formula of recension was applied to texts, classical as well as medieval, for which it was unsuitable, often with grotesque results. Commonly this took the form of choosing on "scientific" (*i.e.,* stemmatic) grounds a "best manuscript" (*codex optimus*) and defending its readings as authoritative even where common sense showed that they could not be authentic. This was the type of editing satirized by A.E. Housman in the brilliant prefaces to his editions of Manilius (1903) and Juvenal (1905) and in many reviews and articles. It flourished chiefly between 1875 and 1900, but the dangers of excessive methodological rigidity had already been foreseen. In 1841 H. Sauppe in his *Epistola Critica ad G. Hermannum* had emphasized the diversity of transmissional situations and the difficulty or actual impossibility of classifying the manuscripts in all cases. In 1843 Lachmann's pupil O. Jahn, in his edition of Persius, had repudiated the strict application of the genealogical method as unsuitable to the tradition of that poet. The most extreme position was taken by E. Schwartz, who in his edition of Eusebius' *Historia ecclesiastica* (1909) denied that "vertically" transmitted texts of Greek books existed at all. The limitations of the stemmatic method have subsequently been stressed in a more temperate fashion by other writers. The modern tendency is to acknowledge the validity of the method in principle while recommending a cautious empiricism in its application. For the editor of a contaminated tradition—and most traditions are probably contaminated—the lesson of recent research is that authoritative evidence may survive even in late and generally corrupt or interpolated sources.

More radical criticism of the method has come from medievalists. In 1913 and again in 1928 the French scholar J. Bédier attacked the stemmatic method because the *stemmata* it produced for medieval texts almost invariably had only two branches. Subsequent investigation has shown that Bédier overrated the inherent improbability of this situation, and it is generally agreed that his criticisms had to do with improper application rather than with the method itself. The point taken by H. Quentin (1922) has already been mentioned: that the method entails argument in a circle, since it relies on the identification of errors at the beginning of a process designed to lead to that very end. This objection, more cogent in theory than in practice, applies with greater force to medieval than to classical texts. The linguistic and stylistic canons of classical Greek and Latin are relatively strict and well defined, whereas the vocabulary, grammar, and usage of many medieval authors (especially when an oral prehistory is in question) is often not certain enough to allow reliable discrimination between variant and error. Classical texts, moreover, have passed through a series of bottlenecks in their history, which have simplified editorial problems by eliminating a high proportion of the evidence (*cf.* the remarks on papyri above). With a few exceptions, such as the commentary of Servius, only one version of each text remains to be reconstructed, whereas many medieval texts are extant in several redactions that cannot be winnowed by the stemmatic method so as to leave only one. Quentin's own method, which depended on the comparison of variants in groups of three, without prejudice as to their correctness, has not been generally adopted. It is immensely laborious and does not in practice possess the objectivity that its inventor claimed for it. Bédier and Quentin have, however, done good service to textual criticism in enjoining caution. The best critics in all fields now agree in rejecting the "logical" (*i.e.,* the illogical) application of any method if the results conflict with common sense, and in stressing the necessity of judging variant readings and forms of a text on their intrinsic merits in the light of the information available.

Bédier and Quentin

Mechanical methods. Quentin also gave a lead to later investigators in calling attention to the possibility of basing recension on the variants themselves, and the more sophisticated methods of Greg (1927), Archibald Hill (1950), Vinton Dearing (1959), and J. Froger (1968) may be seen as a continuation of his work. It has already been suggested that methods of this type, so far as recension is concerned, have been of primarily theoretical interest. But the use of mechanical and computing techniques in this field is in its infancy, and assessment must be provisional. Certain practical applications seem to have proved them-

selves. Mechanical aids to collation have been successfully used in editing Shakespeare and Dryden. Computer storage and analysis of texts can provide information about authorial usage, such as stylistic and metrical patterns, and facilitate the production of concordances. These aids are more relevant to conjectural emendation (as shown by their application to the Dead Sea Scrolls) and the "higher" criticism (*e.g.*, determination of authenticity) than to the recension of texts. The formula or machine that will do the critic's essential work for him still awaits discovery; the best texts are produced by the best scholars, whatever their method or lack of method. Lachmann observed that the establishment of a text according to its tradition is a strictly historical undertaking. Twentieth-century research into the composition and transmission of ancient, medieval, and modern texts has confirmed the truth of his pronouncement. (E.J.Ke.)

BIBLIOGRAPHY

Historiography: C.V. LANGLOIS and C. SEIGNOBOS, *Introduction aux études historiques* (1898; Eng. trans., *Introduction to the Study of History,* 1898); HAROLD TEMPERLEY (ed.), *Selected Essays of J.B. Bury* (1930, reprinted 1964); JAMES T. SHOTWELL, *The History of History* (1939); JAMES W. THOMPSON and BERNARD J. HOLM, *A History of Historical Writing,* 2 vol. (1942); ROBIN G. COLLINGWOOD, *The Idea of History* (1946); HERBERT BUTTERFIELD, *Man on His Past* (1955); JOHN W. MILLER, *The Philosophy of History with Reflections and Aphorisms* (1981); AGNES HELLER, *A Theory of History* (1982).

J.B. BURY, *The Ancient Greek Historians* (1909, reprinted 1958); M.L.W. LAISTNER, *The Greater Roman Historians* (1947, reprinted 1963); MOSES I. FINLEY (ed.), *The Greek Historians* (1959), selected passages in translation with a valuable introduction; MAURICE PLATNAUER (ed.), *Fifty Years of Classical Scholarship* (1954; rev. ed. with appendixes, *Fifty Years (and Twelve) of Classical Scholarship,* 1968), especially chapters 6 by G.T. GRIFFITH and 13 by A.H. MacDONALD; ARNALDO MOMIGLIANO, *Studies in Historiography* (1966), a selection from his vast collection of valuable articles in *Contributo alla storia degli studi classici e del mondo antico,* 5 vol. (1959–69); T.A. DOREY (ed.), *Latin Historians* (1966) and *Latin Biography* (1967); JOHN BARKER, *The Superhistorians: Makers of Our Past* (1982).

GYULA MORAVCSIK, *Byzantinoturcica,* 2nd ed., vol. 1 (1958), not always reliable in its judgments. There is no satisfactory study in English. There is much useful information in the appendixes to J.B. BURY's edition of *The History of the Decline and Fall of the Roman Empire,* 7 vol. (1896–1900, reprinted 1909–14).

THOMAS F. TOUT, "The Study of Mediaeval Chronicles," *Bulletin of the John Rylands Library,* 6:414–438 (1921), reprinted in *The Collected Papers of Thomas Frederick Tout,* 3 vol. (1932–34); REGINALD L. POOLE, *Chronicles and Annals* (1926); M.L.W. LAISTNER, *Thought and Letters in Western Europe, A.D. 500 to 900,* new ed. (1957); CHARLES H. HASKINS, *The Renaissance of the Twelfth Century* (1927); RALPH H.C. DAVIS and JOHN M. WALLACE-HADRILL (eds.), *The Writing of History in the Middle Ages* (1981).

WALLACE K. FERGUSON, *The Renaissance in Historical Thought* (1948); DENYS HAY, "Flavio Biondo and the Middle Ages," *Proceedings of the British Academy,* 45:97–128 (1960); MYRON GILMORE, *Humanists and Jurists* (1963); PAUL O. KRISTELLER, *Eight Philosophers of the Italian Renaissance,* ch. 2 (1964), on Valla; FELIX GILBERT, *Machiavelli and Guicciardini: Politics and History in Sixteenth-Century Florence* (1965); IDA MAIER, *Ange Politien* (1966); L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars: A Guide to the Transmission of Greek and Latin Literature* (1968); ROBERTO WEISS, *The Renaissance Discovery of Classical Antiquity* (1969); E.J. KENNEY, "The Character of Humanist Philology," in R.R. BOLGAR (ed.), *Classical Influences on European Culture, A.D. 500–1500* (1971).

J.G.A. POCOCK, *The Ancient Constitution and the Feudal Law: A Study of English Historical Thought in the Seventeenth Century* (1957), also valuable for France; HERBERT BUTTERFIELD, "The History of Historiography and the History of Science," *Mélanges Alexandre Koyré,* vol. 2 (1964); and "Delays and Paradoxes in the Development of Historiography," in KENNETH BOURNE and D.C. WATT (ed.), *Studies in International History: Essays Presented to W. Norton Medlicott* (1967); GLANMOR WILLIAMS, *Reformation Views of Church History* (1970); DONALD R. KELLEY, *Foundations of Modern Historical Scholarship: Language, Law, and History in the French Renaissance* (1970); MAY MCKISACK, *Medieval History in the Tudor Age* (1971); G. STRAUSS, "The Course of German History: The Lutheran Interpretation," in ANTHONY MOLHO and JOHN A. TEDESCHI (eds.), *Renaissance: Studies in Honor of Hans Baron* (1971).

DAVID C. DOUGLAS, *English Scholars, 1660–1730,* 2nd rev. ed. (1951); MARTIN L. CLARKE, *Greek Studies in England, 1700–1830* (1945); DAVID KNOWLES, "Jean Mabillon," *Journal of Ecclesiastical History,* vol. 10, no. 2 (1959), and *Great Historical Enterprises* (1962); CHRISTOPHER DAWSON, "Edward Gibbon," *Proceedings of the British Academy,* 20:159–180 (1934); and EDWARD GIBBON, *The History of the Decline and Fall of the Roman Empire,* 6 vol. (1776–86, best modern edition by J.B. BURY, *op. cit.*); HUGH TREVOR-ROPER, "The Historical Philosophy of the Enlightenment," in *Studies on Voltaire and the Eighteenth Century,* 27:1667–87 (1963), and "The Idea of the Decline and Fall of the Roman Empire," in *The Age of the Enlightenment* (1967).

J.G.D. CLARK, *Prehistoric Europe: The Economic Basis,* rev. ed. (1962); O.G.S. CRAWFORD, *Archaeology in the Field* (1953); J.G.D. CLARK, *World Prehistory* (1960); CHARLES SAMARAN (ed.), *L'Histoire et ses méthodes,* vol. 11 of the *Encyclopédie de la Pléiade* (1961); VIVIAN H. GALBRAITH, *An Introduction to the Study of History* (1964).

ARTHUR M. SCHLESINGER (ed.), *Historical Scholarship in America* (1932); WILLIAM T. HUTCHINSON (ed.), *The Marcus W. Jernegan Essays in American Historiography* (1937); HUGH H. BELLOT, *American History and American Historians* (1952); DONALD SHEEHAN and HAROLD C. SYRETT (eds.), *Essays in American Historiography: Papers Presented in Honor of Allan Nevins* (1960); JOHN HIGHAM (ed.), *The Reconstruction of American History* (1962); THOMAS J. PRESSLY, *Americans Interpret Their Civil War,* 2nd ed. (1962); RICHARD HOFSTADTER, *The Progressive Historians, Turner, Beard, Parrington* (1968); MARCUS CUNLIFFE and ROBIN W. WINKS (eds.), *Pastmasters: Some Essays on American Historians* (1969); C.L. SONNIGSEN, *The Ambidextrous Historian: Historical Writers and Writing in the American West* (1983).

LORD ACTON, *Historical Essays and Studies* (1907), see especially "German Schools of History"; J.B. BURY, *The Idea of Progress: An Inquiry into Its Origin and Growth* (1920, reprinted 1960); EDMUND WILSON, *To the Finland Station* (1940), especially for French historiography; PIETER GEYL, *Napoleon, voor en tegen in de Franse geschiedschrijving* (1946; Eng. trans., *Napoleon, For and Against,* 1949); *Some Modern Historians of Britain: Essays in Honour of R.L. Schuyler* (1951); G.P. GOOCH, *History and Historians in the Nineteenth Century,* 2nd rev. ed. (1952); FERDINAND SCHEVILL, *Six Historians* (1956), particularly interesting on Ranke; GEORG G. IGGERS, "The Image of Ranke in American and German Historical Thought," in *History and Theory,* 2: 17–123 (1962); and *The German Conception of History* (1968); HENRY E. BELL, *Maitland: A Critical Examination and Assessment* (1965); FREDERICK M. BARNARD, *Herder's Social and Political Thought: From Enlightenment to Nationalism* (1965); JOHN CANNON (ed.), *The Historian at Work* (1980); J.W. BURROW, *A Liberal Descent: Victorian Historians and the English Past* (1981).

EUGENE N. ANDERSON, "Meinecke's *Ideengeschichte* and the Crisis in Historical Thinking," in *Medieval and Historiographical Essays in Honor of James Westfall Thompson* (1938); MARC BLOCH, *Apologie pour l'histoire; ou, Métier d'historien* (1949; Eng. trans., *The Historian's Craft,* 1953); *Architects and Craftsmen in History: Festschrift für Abbott Payson Usher* (1956), biographies of leading historians; JACK H. HEXTER, *Reappraisals in History* (1961); RICHARD PARES, *The Historian's Business, and Other Essays* (1961), particularly valuable for the works of ARNOLD TOYNBEE; FRIEDRICH MEINECKE, *Die Idee der Staatsräson in der neueren Geschichte* (1957; Eng. trans., *Machiavellism: The Doctrine of Raison d'État and Its Place in Modern History,* 1957); DONALD C. WATT (ed.), *Contemporary History in Europe* (1968); LAWRENCE STONE, *The Past and the Present* (1981).

On Russian historiography there is no satisfactory general survey in English. The following can be useful for particular periods or historians: ANATOLE G. MAZOUR, *Modern Russian Historiography,* 2nd ed. (1958); ALEXANDER S. VUCINICH, *Science in Russian Culture: A History to 1860* (1963); and RICHARD PIPES (trans.), *Karamzin's Memoir on Ancient and Modern Russia* (1966); JOHN S. CURTISS (ed.), *Essays in Russian and Soviet History, in Honor of Gerold Tanquary Robinson* (1962), especially on Semevsky; ALAN D. FERGUSON and ALFRED LEVIN (eds.), *Essays in Russian History: A Collection Dedicated to George Vernadsky* (1964); JOHN KEEP and LILIANA BRISBY (eds.), *Contemporary History in the Soviet Mirror* (1964).

FRANZ ROSENTHAL, *A History of Muslim Historiography,* 2nd rev. ed. (1968).

WILLIAM G. BEASLEY and E.G. PULLEYBLANK, *Historians of China and Japan* (1961); CHARLES S. GARDNER, *Chinese Traditional Historiography* (1938, reprinted 1961).

Archaeology: Good general introductions to the aims and methods of archaeology are LEONARD WOOLLEY, *Digging Up the Past* (1930); SIR MORTIMER WHEELER, *Archaeology from the Earth* (1954); and GRAHAME CLARK, *Archaeology and Society,* 3rd rev. ed. (1957). For the history of archaeology and its relation to the development of anthropology, see C.W. CERAM,

Götter, Gräber und Gelehrte (1949; Eng. trans., *Gods, Graves and Scholars,* 1951); G. BIBBY, *The Testimony of the Spade* (1956); and GLYN DANIEL, *A Hundred and Fifty Years of Archaeology* (1974). Anthologies of archaeological writings that relate both to the history of the subject and its present methods are many. The following are recommended: R.F. HEIZER, *The Archaeologist at Work* (1959), and *Man's Discovery of his Past,* 2nd ed. (1970); and JACQUETTA HAWKES, *The World of the Past* (1963). For the development of American archaeology, see G. WILLEY and G. SABLOFF, *The History of American Archaeology* (1973). Special aspects of the development of archaeology are dealt with in D. BROTHWELL and E. HIGGS, *Science in Archaeology,* 2nd ed. (1969); GEORGE F. BASS, *Archaeology Under Water* (1967); W.F. LIBBY, *Radiocarbon Dating,* 2nd ed. (1955); KENNETH HUDSON, *A Social History of Archaeology* (1981); MYRA SHACKLEY, *Environmental Archaeology* (1981); and M.G.L. BAILLIE, *Tree-Ring Dating in Archaeology* (1982).

Bibliography: THEODORE BESTERMAN, *The Beginnings of Systematic Bibliography,* 2nd ed. rev. (1936); A. ESDAILE, *A Student's Manual of Bibliography,* 3rd rev. ed. (1954); L.N. MALCLES, *Les Sources du travail bibliographique,* 3 vol. (1950–58); GEORG SCHNEIDER, *Handbuch der Bibliographie,* 4th ed. (1930, reprinted 1969); J.H. SHERA and M.E. EGAN (eds.), *Bibliographic Organization* (1951); RONALD STAVELEY, I.C. MCILWAINE, and JOHN H. ST. MCILWAINE, *Introduction to Subject Study* (1967); R. STOKES, *The Function of Bibliography* (1969); CURT F. BUHLER *et al., Standards of Bibliographical Description* (1949); J.D. COWLEY, *Bibliographical Description and Cataloguing* (1939); R.B. MCKERROW, *An Introduction to Bibliography for Literary Students* (1927); F.T. BOWERS, *Bibliography and Textual Criticism* (1964) and *Principles of Bibliographical Description* (1949); O.M. BRACK, JR. and W. BARNES (eds.), *Bibliography and Textual Criticism: English and American Literature, 1700 to the Present* (1969); ROBERT B. HARMON, *Elements of Bibliography* (1981).

Chronology: JAMES C. MACDONALD, *Chronologies and Calendars* (1897); ALFRED E. STAMP, *Methods of Chronology* (1933); R.L. POOLE, *Studies in Chronology and History,* collected and ed. by A.L. POOLE (1934, reprinted 1969).

On the astronomical basis of Chinese calendrical systems, see JOSEPH NEEDHAM, *Science and Civilisation in China,* vol. 3, pp. 390–408 (1959). A standard reference for conversion between Chinese and Western calendars is MATHIAS TCHANG, *Synchronismes chinois* (1905).

ROBERT SEWELL and S.B. DIKSHITA, *The Indian Calendar* (1896), describes the various systems of calendars in India, with tables of concordance and a useful index; ROBERT SEWELL, *The Siddhantas and the Indian Calendar* (1924), a study of the Hindu astronomical system as a basis for the traditional calendar; SWAMIKANNU PILLAI, *An Indian Ephemeris,* 6 vol. (1922), tables of concordance of Hindu, Muslim, and modern calendars; JEAN FILLIOZAT, "Notions de chronologie," in LOUIS RENOU and JEAN FILLIOZAT, *Inde classique,* vol. 2 (1953), a general summary and list of different eras used in India.

RICHARD A. PARKER, *The Calendars of Ancient Egypt* (1950), a basic work providing valid solutions to most of the problems of Egyptian chronology; "Lunar Dates of Thutmose III and Ramesses II," *J. Near Eastern Stud.,* 16: 39–43 (1957), an analysis of the more important lunar dates of the New Kingdom; M.B. ROWTON, "Manetho's Date for Ramesses II," *J. Egyptian Archaeol.,* 34:57–74 (1948), and "Comparative Chronology at the Time of Dynasty XIX," *J. Near Eastern Stud.,* 19:15–22 (1960), two articles that deal with the date for the accession of Ramses II (1290 or 1304 BC); ERIK HORNUNG, *Untersuchungen zur Chronologie und Geschichte des Neuen Reiches* (1964), the most up-to-date synthesis of the chronological problems of the New Kingdom; I.E.S. EDWARDS, "Absolute Dating from Egyptian Records and Comparison with Carbon-14 Dating," *Phil. Trans. R. Soc.,* Series A, 269:11–18 (1970), includes a complete list of all radiocarbon dates taken from Egyptian samples.

The most complete general work on the chronology of Western Asia, including Mesopotamia, is P.E. VAN DER MEER, *The Chronology of Ancient Western Asia and Egypt,* 2nd rev. ed. (1955); however, it should be used with caution. For the general chronology of ancient Western Asia, with special reference to Mesopotamia, see M.B. ROWTON in *The Cambridge Ancient History,* 3rd ed., vol. 1, pp. 193–237 (1970), and the extensive literature quoted there (up to 1959). For a discussion of the gap in the 2nd millennium BC, see A. GOETZE, "The Kassites and Near Eastern Chronology," *J. Cuneiform Stud.,* 18:97–101 (1964); and B.L. VAN DER WAERDEN, *Die Anfänge der Astronomie,* pp. 28–49 (1966). For the chronology of the late 2nd millennium BC and the chronological gap in the 1st millennium BC, see J.A. BRINKMAN, *A Political History of Post-Kassite Babylonia, 1158–722 B.C.,* pp. 37–85 (1968).

B. RATNER, *Seder Olam Rabba: Die grosse Weltchronik* (1897), the authoritative edition, carefully annotated and with a detailed introduction, of the oldest rabbinic chronology extending from the earliest records to the first century of the current era; E. MAHLER, *Handbuch der jüdischen Chronologie* (1916), the only comprehensive work on Jewish chronology, covering all aspects and summarizing most previous works on the subject; S. ZEITLIN, *Megillat Taanit As a Source for Jewish Chronology and History in the Hellenistic and Roman Periods* (1922), removes all discrepancy in I and II Maccabees' and Josephus' chronological data in the Hasmonean phase of Jewish history, thus reinvesting their statements with historical significance and authority.

ALAN E. SAMUEL, *Greek and Roman Chronology: Calendars and Years in Classical Antiquity* (1972), a fundamental work; HENRY FYNES CLINTON, *Fasti Hellenici,* 3 vol. (1834), still useful in providing the sources for the framework of Greek chronology; W. KUBITSCHEK, *Grundriss der antiken Zeitrechnung* (1928), the standard handbook but unreliable; A.E. SAMUEL, *Ptolemaic Chronology* (1962); E.J. BICKERMAN, *Chronology of the Ancient World* (1968), a summary survey to be used judiciously; W. DEN BOER, *Laconian Studies,* pt. 1, "The Struggle for the Chronological Pattern" (1954), on ancient attempts to systematize archaic chronology; A.G. WOODHEAD, *The Greeks in the West,* pp. 69–72 (1962), on dates in the archaic period.

E.J. BICKERMAN, *Chronology of the Ancient World* (1968), an excellent short manual, unfortunately marred by a number of factual errors; AGNES KIRSOPP MICHELS, *The Calendar of the Roman Republic* (1967), contains much valuable information on chronology as well as the calendar—probably the best book ever written on the Roman calendar; F.K. GINZEL, *Handbuch der mathematischen und technischen Chronologie das Zeitrechnungswesen der Völker,* 3 vol. (1906–14), the standard work of reference on its subject.

F.K. GINZEL, *Handbuch der mathematischen und technischen Chronologie das Zeitrechnungswesen der Völker,* 3 vol. (1906–14), still the fundamental work; JACK FINEGAN, *Handbook of Bible Chronology* (1964), the best for its subject; E.J. BICKERMAN, *Chronology of the Ancient World* (1968), very thorough and complete.

AL-BIRUNI, *The Chronology of Ancient Nations,* Eng. trans. by C. EDWARD SACHAU (1879), useful source material on the Muslim knowledge of chronology down to the 10th century AD; see especially pp. 16–82. For actual tables of Muslim chronological information and comments, the following may be consulted: E. LACOINE, *Tables de concordance des dates des calendriers . . .* (1891); C.L. IDELER, *Handbuch der mathematischen und technischen chronologie,* 2 vol. (1825–26); S. LANE-POOLE, *The Mohammedan Dynasties* (1894); and C.E. BOSWORTH, *The Islamic Dynasties* (1967). See also articles on calendar, chronology, and "Hidjra" in the *Encyclopaedia of Islam.* An extensive collection of conversion tables is given in FRANK PARISE (ed.), *The Book of Calendars* (1982).

S.G. MORLEY, "An Introduction to the study of the Maya Hieroglyphs," *Bull. U.S. Bur. Am. Ethnol.,* no. 57 (1915, reprinted 1968), a clear exposition of Maya chronology; J. ERIC S. THOMPSON, *Maya Chronology: The Correlation Question* (1935) and *Maya Hieroglyphic Writing: Introduction* (1950, reprinted 1971); H.J. SPINDEN, *The Reduction of Maya Dates* (1924), a presentation of earlier correlation; L. SATTERTHWAITE and E.K. RALPH, "New Radiocarbon Dates and the Maya Correlation Problem," *Am. Antiq.,* 26:165–184 (1960).

Diplomatics: HARRY BRESSLAU, *Handbuch der Urkundenlehre für Deutschland und Italien,* 2nd ed., vol. 1 (1912), vol. 2, pt. 1 (1914), vol. 2, pt. 2, ed. by H.W. KLEWITZ (1931, all reprinted 1958, with separate index), although somewhat out of date, still the best handbook by far for Germany and Italy; OSWALD REDLICH, *Die Privaturkunden des Mittelalters* (1911, reprinted 1967), an excellent work on the private documents of the Middle Ages; LEO SANTIFALLER, *Beiträge zur Geschichte der Beschreibstoffe im Mittelalter,* pt. 1 (1953), the most up-to-date account of the history of writing materials during the Middle Ages; FRANZ DOLGER and JOHANNES KARAYANNOPULOS, *Byzantinische Urkundenlehre die Kaiserurkunden* (1968), the only study of the documents of the Byzantine emperors; C.R. CHENEY, *The Study of the Medieval Papal Chancery* (1966), an excellent general survey of modern research; R.L. POOLE, *Lectures on the History of the Papal Chancery down to the Time of Innocent III* (1915), the best book in English on this subject, though now out of date; PETER HERDE, *Beiträge zum päpstlichen Kanzlei- und Urkundenwesen im 13. Jahrhundert,* 2nd ed. (1967); and *Audientia litterarum contradictarum,* 2 vol. (1970), on papal letters of justice; WILHELM ERBEN, *Die Kaiser- und Königsurkunden des Mittelalters in Deutschland, Frankreich und Italien* (1907, reprinted 1967), an excellent supplement to Bresslau on the imperial and royal documents of the Middle Ages in Germany, France, and Italy; GEORGES TESSIER, *Diplomatique royale française* (1962), the best and most up-to-date handbook on the royal French diplomatic; ALAIN DE BOUARD, *Manuel de diplomatique française et pontificale* (1929), a handbook on French and papal diplomatic;

F.M. STENTON, *The Latin Charters of the Anglo-Saxon Period* (1955), a good brief survey on research since the 18th century; P.H. SAWYER, *Anglo-Saxon Charters* (1968), the most up-to-date annotated list and bibliography; R.C. VAN CAENEGEM, *Royal Writs in England from the Conquest to Glanvill* (1959), the best book on writs; V.H. GALBRAITH, *An Introduction to the Use of the Public Records* (1934) and *Studies in the Public Records* (1948), two works especially useful for the later periods; C.R. CHENEY, *The Records of Medieval England* (1956) and *English Bishops' Chanceries 1100–1250* (1950); T.F. TOUT, *Chapters in the Administrative History of Medieval England,* 6 vol. (1920–33), the basic work on the subject, including the chancery; HARRY BRESSLAU, "International Beziehungen im Urkundenwesen des Mittelalters," *Archiv für Urkundenforschung,* 6:19–76 (1918), the only survey on international relations in the documentary system of the Middle Ages; HORST ENZENSBERGER, *Beiträge zum Kanzlei- und Urkundenwesen der normannischen Herrscher Unteritaliens und Siziliens* (1971), the most up-to-date account of the chancery and documents of the Norman rulers in southern Italy and Sicily; H.O. MEISNER, *Archivalienkunde vom 16. Jahrhundert bis 1918* (1969), the best study of records from the 16th century through 1918 for central Europe, especially Germany; CHARLES CARTER, *The Western European Powers, 1500–1700* (1971), a very useful survey on archives and their records. All the cited works have comprehensive bibliographies, with some of them also containing bibliographies of editions and facsimiles of documents. H. THOMAS HICKERSON, *Archives and Manuscripts: An Introduction to Automated Access* (1981), discusses specific concern of the contemporary archives business.

Epigraphy: J.B. PRITCHARD (ed.), *Ancient Near Eastern Texts Relating to the Old Testament,* 3rd ed. (1969), gives translations of inscriptional material, especially from Egypt, Mesopotamia, Palestine, and Asia Minor. J. FRIEDRICH, *Entzifferung verschollener Schriften und Sprachen* (1954; Eng. trans., *Extinct Languages,* 1957), deals especially with the decipherment of inscriptions, while his *Kleinasiatische Sprachdenkmäler* (1932), transliterates a variety of epichoric texts from Asia Minor. R.G. KENT, *Old Persian,* 2nd ed. rev. (1953), is a comprehensive work including grammar, texts, and lexicon. W. HINZ, *Altiranische Funde und Forschungen* (1969), includes a number of important recent inscriptional finds from ancient Iran. J. BLOCH, *Les Inscriptions d'Aśoka* (1950), gives texts and translations of all Aśokan epigraphs; whereas R.B. PANDEY, *Historical and Literary Inscriptions* (1962), provides a transliterated selection of later Indic epigraphy as well. L. DEROY, *Initiation à l'épigraphie mycénienne* (1962), provides a comprehensive survey of the study of Mycenaean tablets. E.S. ROBERTS, *An Introduction to Greek Epigraphy,* 2 vol. (1887–1905), remains the classic compendium of its kind; while more recent, shorter surveys of its subject are found in G. KLAFFENBACH, *Griechische Epigraphik* (1957); and A.G. WOODHEAD, *The Study of Greek Inscriptions* (1959). A more elaborate recent treatment is that of M. GUARDUCCI, *Epigrafia greca,* 2 vol. (1967–69). The Cretan laws have been sumptuously edited and commented in R.F. WILLETTS, "The Law Code of Gortyn," *Kadmos,* suppl. 1 (1967). The old standard works on Latin inscriptions, J.E. SANDYS, *Latin Epigraphy,* 2nd ed. rev. (1927); and R.L.V. CAGNAT, *Cours d'épigraphie latine,* 3rd ed. (1898), are updated only in such summary surveys as R. BLOCH, *L'épigraphie latine,* 3rd ed. (1964). The elaborate edition of the principal Umbrian inscription by J.W. POULTNEY, *The Bronze Tables of Iguvium* (1959), gives text, translation, grammar, extensive commentary, and facsimiles of the tables themselves. R.W.V. ELLIOTT, *Runes: An Introduction* (1959), is especially thorough on the side of British runes; while W. KRAUSE, *Runen* (1970), is a compact but comprehensive survey of the entire field of runology.

Genealogy: For an outline of the sources and methods of procedure, see L.G. PINE, *Heraldry and Genealogy,* 4th ed. (1974). Also useful is A.J. WILLIS, *Genealogy for Beginners,* 3rd. ed. (1976). J. UNETT, *Making a Pedigree,* 2nd ed. (1961), contains information on medieval records. G. HAMILTON EDWARDS, *In Search of Ancestry* (1966), is a larger work with more detail, especially on naval and military sources. On American genealogy, see G.H. DOANE and J.B. BELL, *Searching for Your Ancestors: The How and Why of Genealogy,* 5th ed. (1980); and J.S. SWEET, *Genealogy and Local History: An Archival and Bibliographical Guide,* 2nd rev. ed. (1959). See also L.G. PINE, *American Origins* (1960, reprinted 1980), on sources for genealogical research in Europe, and *The Genealogist's Encyclopedia* (1969). A good starting point in genealogical quests is P. WILLIAM FILBY, *American and British Genealogy and Heraldry: A Selected List of Books,* 3rd ed. (1983).

Paleography: The most comprehensive work, certainly in English, is still E. MAUNDE THOMPSON, *An Introduction to Greek and Latin Palaeography* (1912). Also valuable, and in print, is his shorter version, *Handbook of Greek and Latin Palaeography* (1893, reprinted 1966). For Greek paleography, see B.A. VAN GRONINGEN, *Short Manual of Greek Palaeography,* 3rd rev. ed. (1963); and C.H. ROBERTS, *Greek Literary Hands, 350 B.C.–A.D. 400* (1956). The article "Handwriting" in C.G. CRUMP and E.F. JACOB (eds.), *The Legacy of the Middle Ages* (1926), is an important account of Latin book hands. Essential for the study of abbreviations in medieval Europe is ADRIANO CAPPELLI, *Lexicon abbreviaturarum: dizionario di abbreviature latine ed italiane,* 6th ed. (1961). A standard textbook for medieval English cursive hands is CHARLES JOHNSON and HILARY JENKINSON, *English Court Hand A.D. 1066 to 1500* (1915), continued in Jenkinson's *Later Court Hands in England from the Fifteenth to the Seventeenth Century* (1927). L.C. HECTOR, *The Handwriting of English Documents,* 2nd ed. (1966), contains a valuable introduction to the paleography of English administrative manuscripts. Very many classical, biblical, and liturgical texts have now been published in facsimile, often with colour plates. Large collections in facsimile comprise E.A. LOWE, *Codices latini antiquiores,* 12 vol. (1934–71); CHARLES SAMARAN and ROBERT MARICHAL, *Catalogue des manuscrits en écriture latine, portant des indications de date* (1959–); BERTRAM COLGRAVE (ed.), *Early English Manuscripts in Facsimile* (1951–); and the "Oxford Palaeographical Handbooks," a series designed to deal with various aspects of the subject, such as C.H. ROBERTS (see above); C.E. WRIGHT, *English Vernacular Hands from the Twelfth to the Fifteenth Centuries* (1960); M.B. PARKES, *English Cursive Book Hands, 1250–1500* (1969); and T.A.M. BISHOP, *English Caroline Minuscule* (1971). See also MARC DROGIN, *Medieval Calligraphy, Its History and Technique* (1980); and DONALD JACKSON, *The Story of Writing* (1981).

Sigillography: The best general account is SIR HILARY JENKINSON, *Guide to Seals in the Public Record Office* (1954). JOSEPH H. ROMAN, *Manuel de sigillographie française* (1912), is a longer study mainly on French seals. The basic catalog for European seals and impressions is WALTER DE GRAY BIRCH, *Catalogue of Seals in the Department of Manuscripts in the British Museum,* 6 vol. (1887–1900). Seal matrices are discussed by ALEC B. TONNOCHY in the *Catalogue of British Seal-Dies in the British Museum* (1952). ALFRED B. WYON, *The Great Seals of England from the Earliest Period to the Present Time* (1887), provides a systematic and well-illustrated account of English Great Seals. A full discussion of Classical seal use, with bibliography and occasional reference to Near Eastern sources, is given in the article "Signum," by WENGER in *Pauly-Wissowa Real-Encyclopädie,* vol. 2, col. 2361–2448 (1923); and details of use in the Aegean and Greece are conveniently given in JOHN BOARDMAN, *Greek Gems and Finger Rings* (1970). No adequate study exists of the uses of seals in ancient western Asia. See at present, however, ELENA CASSIN, "Le Sceau: un fait de civilisation dans la Mésopotamie ancienne," *Annales,* pp. 742–751 (1960); and M.I. ROSTOVTSEFF, "Seleucid Babylonia: Bullae and Seals of Clay with Greek Inscriptions," *Yale Classical Studies* 3:1–114 (1932). A very full bibliography of ancient Eastern seal publications from which information on use may be gleaned is given by HANS H. VON DER OSTEN in *Ancient Oriental Seals in the Collection of Mr. Edward T. Newell,* pp. 168–190 (1934); and in *Altorientalische Siegelsteine der Sammlung Hans Silvius von Aulock,* pp. 156–219 (1957), but no more recent convenient bibliography exists; systematic study of use must depend mainly on the examination of sealed texts. For Egypt, P.E. NEWBERRY, *Scarab-Shaped Seals* (1907), provides a brief account, now much in need of revision. Books on Chinese and Japanese seals deal mainly with those of painters, calligraphers, and collectors. ROBERT H. VAN GULIK, *Chinese Pictorial Art As Viewed by the Connoisseur* (1958), has the most illuminating discussion of the artist's use of seals. See also VICTORIA CONTAG and WANG CHI-CH'IEN, *Seals of Chinese Painters and Collectors of the Ming and Ch'ing Periods,* rev. ed. (1966); and CH'EN CHIH-MAI, *Chinese Calligraphers and Their Art* (1966).

Textual criticism: For oral transmission the fundamental work is H.M. and N.K. CHADWICK, *The Growth of Literature,* 3 vol. (1932–40, reprinted 1969). Classical, biblical, and medieval texts are all covered (but somewhat unevenly) by H. HUNGER *et al., Geschichte der Textüberlieferung der antiken und mittelalterlichen Literatur,* 2 vol. (1961–64). For classical texts in general the best introduction is L.D. REYNOLDS and N.G. WILSON, *Scribes and Scholars* (1968); more specialized are A.C. CLARK, *The Descent of Manuscripts* (1918); and A. DAIN, *Les manuscrits,* 2nd rev. ed. (1964). For special areas, see B.A. VAN GRONINGEN, *Traité d'histoire et de critique des textes grecs* (1963); R.D. DAWE, *The Collation and Investigation of Manuscripts of Aeschylus* (1964); R. RENEHAN, *Greek Textual Criticism: A Reader* (1969); L. HAVET, *Manuel de critique verbale appliquée aux textes latins* (1911), an exhaustive catalogue raisonné of scribal error; E.G. TURNER, *Greek Papyri: An Introduction* (1968), important for the history of Greek texts in antiquity. For biblical texts, in addition to Hunger (*op. cit.*), the standard work on the New Testament is B.M. METZGER,

The Text of the New Testament, 2nd ed. (1968). For patristic texts an excellent case study raising important general issues is M. BEVENOT, *The Tradition of Manuscripts: A Study in the Transmission of St. Cyprion's Treatises* (1961). For medieval texts a good exposition of typical problems may be found in D'A.S. AVALLE, "Die Liederhandschriften und die Textkritik," in Hunger (*op. cit.*), 2:273–290; see also J.A. ASHER, "Truth and Fiction: The Text of Medieval Manuscripts," *Aumla,* 25:6–16 (1966). For printed books the standard work is F. BOWERS, *Bibliography and Textual Criticism* (1964); while a useful collection of essays may be found in O.M. BRACK and W. BARNES (eds.), *Bibliography and Textual Criticism: English and American Literature, 1700 to the Present* (1969).

For the history of the stemmatic method and the contribution of Lachmann, S. TIMPANARO, *La genesi del metodo del Lachmann* (1963), is definitive. For its application, P. MAAS, *Textual Criticism* (Eng. trans. 1958), an austere theoretical exposition; and G. PASQUALI, *Storia della tradizione e critica del testo,* 2nd ed. (1952), brilliant but discursive, are complementary and fundamental. The reaction to the stemmatic method may be studied in H. QUENTIN, *Mémoire sur l'établissement du texte de la Vulgate* (1922), *Essais de critique textuelle–ecdotique* (1926); J. BURKE SEVERS, "Quentin's Theory of Textual Criticism," *English Inst. Annual 1941,* pp. 65–93 (1942); J. BEDIER, "La tradition manuscrite du Lai de L'Ombre: Réflexions sur l'art d'éditer les anciens textes," *Romania,* 54:161–96, 321–56 (1928). A. CASTELLANI, *Bédier avait-il raison?* (1957); W.W. GREG, *The Calculus of Variants* (1927); A.A. HILL, "Some Postulates for Distributional Study of Texts," *Stud. in Biblphy.,* 3:63–95 (1950). For further discussion of these developments, see W.P. SHEPARD, "Recent Theories of Textual Criticism," *Mod. Philol.,* 28:129–141 (1930); J. ANDRIEU, "Principes et recherches en critique textuelle," in *Mémorial des études latines . . . J. Marouzeau,* pp. 458–74 (1943); E.B. HAM, "Textual Criticism and Common Sense," *Romance Philol.,* 12:198–215 (1959). For mechanized techniques, see V.A. DEARING, *A Manual of Textual Analysis* (1959); J. FROGER, *La critique des textes et son automatisation* (1968). For the taxonomic approach, see J.G. GRIFFITH, "A Taxonomic Study of the Manuscript Tradition of Juvenal," *Mus. Helv.,* 25:101–138 (1968); "Numerical Taxonomy and Some Primary Manuscripts of the Gospels," *J. Theol. Stud.,* 20:390–406 (1969).

For classical texts on textual criticism the standard works are still: O. STAHLIN, *Editionstechnik,* 2nd ed. (1914); and A. DELATTE and A. SEVERYNS, *Emploi des signes critiques* (1938). For medieval texts, see A. DONDAINE, "Abbréviations latines et signes raccommandés pour l'apparat critique des éditions des textes médiévaux," *Bull. de la soc. int. pour l'étude de la philosophie méd.,* 2:142–149 (1960); and for papyri, see Turner (*op. cit.*); and H.C. YOUTIE, *The Textual Criticism of Documentary Papyri: Prolegomena* (1958).

Many of the works cited above include discussion of general principles of textual criticism. In addition, the following works may be mentioned: H. KANTOROWICZ, *Einführung in die Textkritik* (1921), sound emphasis on the roles of probability and common sense; A.E. HOUSMAN, "The Application of Thought to Textual Criticism," *Proc. Class. Assoc.,* 18:67–84 (1922), a brilliant polemic against hard-and-fast "rules" of criticism; L. BIELER, "The Grammarian's Craft," *Folia,* 10:3–42 (1956), a useful summary discussion of modern developments; and A.H. MCDONALD, "Textual Criticism," in *Oxford Classical Dictionary,* 2nd ed. (1970), a sound exposition for classical students. JEROME J. MCGANN, *A Critique of Modern Textual Criticism* (1983), is a study of the role of non-literary contexts in textual criticism.

(E.B.Fr./G.E.D./F.C.F./W.S.-cg./J.L.A.F./W.H./M.B.R./E.J.Wi./A.G.W./N.A.Z./F.Th./Pe.He./J.Pl./L.G.P./W.G.U./J.Ch./T. C.M./W.W./E.J.Ke./Ed.)

Hitler

The dictator of Nazi Germany, Adolf Hitler, was born on April 20, 1889, at Braunau am Inn, Austria-Hungary. His father, Alois (born 1837), was illegitimate and for a time bore his mother's name, Schicklgruber, but by 1876 he had established his claim to the surname Hitler. Adolf never used any other name, and the name Schicklgruber was revived only by his political opponents in Germany and Austria in the 1930s.

Heinrich Hoffmann, Munich



Hitler reviewing troops on the eastern front, 1939.

Early life. Adolf Hitler spent most of his childhood in the neighbourhood of Linz, the capital of Upper Austria, after his father's retirement from the Habsburg customs service. Alois Hitler died in 1903 but left an adequate pension and savings to support his wife and children. Adolf received a secondary education and, although he had a poor record at school and failed to secure the usual certificate, did not leave until he was 16 (1905). There followed two idle years in Linz, when he indulged in grandiose dreams of becoming an artist without taking any steps to prepare for earning his living. His mother was overindulgent to her willful son, and even after her death in 1908 he continued to draw a small allowance with which at first he maintained himself in Vienna. His ambition was to become an art student, but he twice failed to secure entry to the Academy of Fine Arts. For some years he lived a lonely and isolated life, earning a precarious livelihood by painting postcards and advertisements and drifting from one municipal lodging house to another.

Hitler already showed traits that characterized his later life: inability to establish ordinary human relationships; intolerance and hatred both of the established bourgeois world and of non-German peoples, especially the Jews; a tendency toward passionate, denunciatory outbursts; readiness to live in a world of fantasy and so to escape his poverty and failure.

In 1913 Hitler moved to Munich. Temporarily recalled to Austria to be examined for military service (February 1914), he was rejected as unfit; but when World War I broke out he volunteered for the German army and joined the 16th Bavarian Reserve Infantry Regiment. He served throughout the war, was wounded in October 1916, and was gassed two years later. He was still hospitalized when the war ended. Except when hospitalized, he was continuously in the front line as a headquarters runner; his bravery in action was rewarded with the Iron Cross, Second Class, in December 1914, and the Iron Cross, First Class (a rare decoration for a corporal), in August 1918. He greeted the war with enthusiasm, as a great relief from the frustration and aimlessness of his civilian life. He found comradeship, discipline, and participation in conflict intensely satisfying and was confirmed in his belief in authoritarianism, inequality, and the heroic virtues of war.

Military service in World War I

Rise to power. Discharged from the hospital in the atmosphere of confusion that followed the German defeat, Hitler determined to take up political work in order to destroy a peace settlement that he denounced as intolerable. He remained on the roster of his regiment until April 1920 and as an army political agent joined the tiny German Workers' Party in Munich (September 1919).

In 1920 he was put in charge of the party's propaganda and left the army to devote his time to building up the party, which in that year was renamed the Nationalsozialistische Deutsche Arbeiterpartei (of which Nazi was an abbreviation). Conditions were ripe for the development of such a party. Resentment at the loss of the war and the peace terms added to economic chaos brought widespread discontent. This was sharpened in Bavaria, where Hitler lived throughout the 1920s, by traditional separatism and dislike of the republican government in Berlin. In March 1920 a coup d'etat by the army established a strong right-wing government. Munich became the gathering place for dissatisfied former servicemen and members of the Freikorps, which had been organized in 1918–19 from units of the German army unwilling to return to civilian life, and for political plotters against the republic. Many of these joined the Nazi Party. Foremost among them was Ernst Röhm, a member of the staff of the district army command, who had actually joined the German Workers' Party before Hitler and who was of great help in furthering his schemes for developing it into an instrument of power. It was he who recruited the "strong arm" squads used by Hitler to protect party meetings, to attack Socialists and Communists, and to exploit violence for the impression of strength it gave. In 1921 these were formally organized under Röhm into a private party army, the SA (Sturmabteilung). Röhm was also able to ensure the protection of the Bavarian government, which depended on the local army command for the maintenance of order and which tacitly accepted his breaches of law and his policy of intimidation.

Although conditions were thus favourable to the growth of the party, only Hitler was sufficiently astute to take full advantage of them. When he joined the party he found it small, ineffective, committed to a program of nationalist and socialist principles but uncertain of its aims and divided in its leadership. He accepted its program but regarded it only as a means to an end—political power. His propaganda methods and his personal arrogance caused friction with the other members of the committee, which was resolved when Hitler countered their attempts to curb his freedom by offering his resignation. Aware that the future of the party depended on his power to organize publicity and to acquire funds, they were forced to give in, and in July 1921 he became president with unlimited powers. From the first he set out to create a mass movement, whose mystique and force would be sufficient to bind its members in loyalty to him. He engaged in unrelenting propaganda through the party newspaper, the *Völkischer*

Beobachter ("Popular Observer," acquired in 1920), and through a succession of meetings, rapidly growing from audiences of a handful to thousands, where he developed his unique talent for magnetism and mass leadership. At the same time, he gathered around him several of the Nazi leaders who later became infamous—Alfred Rosenberg, Rudolf Hess, Hermann Göring, and Julius Streicher.

The Munich *Putsch*

The climax in this rapid growth of the Nazi party in Bavaria came in an attempt to seize power in the Munich (Beer Hall) *Putsch* of November 1923, when Hitler and Gen. Erich Ludendorff took advantage of the prevailing lawlessness and opposition to the Weimar Republic to force the leaders of the *Land* government and the local Reichswehr commander to proclaim a national revolution. When released, however, they rescinded the proclamation. When placed on trial, Hitler, although his part in the *Putsch* had been far from glorious, characteristically took advantage of the immense publicity afforded to him. He also drew a vital lesson from the *Putsch*—that the movement must achieve power by legal means. He was sentenced to prison for five years. but served only nine months, and that in comfort at Landsberg. He used the time to prepare the first volume of *Mein Kampf.*

Mein Kampf

Hitler's ideas included little that cannot be traced to earlier writers or to the commonly accepted shibboleths of Viennese right-wing radicalism in his youth. He regarded inequality between races and individuals as part of an unchangeable natural order and exalted the "Aryan race" as the sole creative element of mankind. The natural unit of mankind was the *Volk,* of which the German was the greatest; and the state only existed to serve the *Volk*—a mission that the Weimar Republic betrayed. All morality and truth was judged by this criterion: whether it was in accordance with the interest and preservation of the *Volk.* For this reason democratic government stood doubly condemned. It assumed an equality within the *Volk* that did not in fact exist, and it supposed that what was in the interests of the *Volk* could be decided by discussion and voting. In fact the unity of the *Volk* found its incarnation in the *Führer,* endowed with absolute authority. Below the *Führer* the party (which Hitler often called the "movement" to distinguish it from democratic parties) was drawn from the best elements of the *Volk* and was in turn its safeguard.

The greatest enemy of Nazism was not, in Hitler's view, liberal democracy, which was already on the verge of collapse. It was rather the rival *Weltanschauung,* Marxism (which for him embraced Social Democracy as well as Communism), with its insistence on internationalism and class conflict. Behind Marxism he saw the greatest enemy of all, the Jew, who was for Hitler the very incarnation of evil, a mythical figure into which he projected all that he feared and hated.

During Hitler's absence in prison the Nazi Party disintegrated through internal dissension. In the task of reconstruction after his release, he faced difficulties that had not existed before 1923. Economic stability had been achieved by currency reform and the Dawes Plan; the republic had become more respectable. Hitler was forbidden to make speeches, first in Bavaria, then in many other German states (these prohibitions remained in force until 1927–28). Nevertheless, the party grew slowly in numbers, and in 1926 Hitler successfully established his position against Gregor Strasser, who had built up a rival Nazi movement in north Germany.

Alliance with Alfred Hugenberg

The slump of 1929 opened a new period of economic and political instability. Hitler made an alliance with the Nationalist Alfred Hugenberg in a campaign against the Young Plan. Through it Hitler was able for the first time to reach a nationwide audience with the help of Hugenberg's Nationalist Party organization and the newspapers it controlled. It also enabled him to commend himself as a gifted agitator to the magnates of business and industry who controlled political funds and were anxious to use them to establish a strong right-wing, anti-working-class government. The subsidies he received from the industrialists placed his party on a secure financial footing and enabled him to make effective his emotional appeal to the lower middle class and the unemployed, based on the proclamation of his faith that Germany would awaken from its sufferings to reassert its natural greatness. Like his later intrigues with the conservatives, Hitler's dealings with Hugenberg and the industrialists exemplify his skill in using those who sought to use him.

Mass agitation and unremitting propaganda, set against the failure of the government to achieve any success in internal or external affairs, produced a steadily mounting electoral strength for the Nazis, who became the second largest party in the country, with more than 6,000,000 votes at the 1930 election. Hitler opposed Hindenburg in the presidential election of 1932, capturing 36.8 percent of the votes on the second ballot.

Placed in a very strong position by his unprecedented mass following, he took part in a series of intrigues for the favour of the aging president in which the other principal participants were Franz von Papen, Gen. Kurt von Schleicher, Otto Meissner, and Hindenburg's son, Oskar. In spite of a decline in the party's votes in November 1932, he held to the chancellorship as the only office he would accept, and this by constitutional, not revolutionary, methods. Throughout, he showed a unique ability to exploit conditions favourable to success. He created the Hitler myth; he propagated it by every device of mass agitation and with an actor's ability to be absorbed in the role that he created for himself. Yet all the time he remained a shrewd and calculating politician, aware of the weaknesses of his own position, perceiving more quickly than anyone else how a situation could best be turned to his own advantage. In January 1933 he reaped his reward when Hindenburg invited him to be chancellor of Germany, and he took office with the support of Papen and Hugenberg and with Field Marshal Werner von Blomberg as minister of defense.

Hitler's personal life had grown more relaxed and stable with the added comfort that accompanied the success of the party. After his release from prison, he went to live on the Obersalzberg, near Berchtesgaden. His income at this time was derived in a haphazard manner from party funds and from writing in nationalist newspapers. When he became chancellor he accepted the material comforts that followed but remained independent of them. He was indifferent to clothes and food, never smoking or drinking tea, coffee, or alcohol. He continued, even as *Führer,* to rebel against routine or regular work—a characteristic that he ascribed to his artistic temperament.

When he went to live at Berchtesgaden, his half sister Angela Raubal and her two daughters accompanied him. Hitler became devoted to one of them, Geli, but his possessive jealousy drove her to suicide in September 1931. For weeks Hitler was inconsolable. Later Eva Braun, a shop assistant from Munich, became his mistress. Hitler rarely allowed her to come to Berlin or appear in public with him and would not consider marriage on the grounds that it would hamper his career. Eva was a warmhearted girl with no intellectual ability. Her great virtue in Hitler's eyes was her unquestioning loyalty, and in recognition of this he made her his legal wife at the end of his life.

Eva Braun

Dictator, 1933–39. Once in power, Hitler proceeded to establish an absolute dictatorship. He secured the President's assent for new elections on the grounds that a majority in the Reichstag could not, after all, be obtained. The Reichstag fire, on the night of February 27, 1933 (apparently the work of a Dutch Communist, Marinus van der Lubbe), provided an excuse for a decree overriding all guarantees of freedom and for an intensified campaign of violence. In these conditions, when the elections were held (March 5), the Nazis polled 43.9 percent of the votes. The Reichstag assembled in the Potsdam Garrison Church, a theatrical gathering designed by Hitler to show the unity of his own movement with the old conservative Germany, represented by Hindenburg. Two days later an enabling bill, giving full powers to Hitler, was passed in the Reichstag by the combined votes of Nazi, Nationalist, and Centre party deputies (March 23, 1933).

Thus far successful, Hitler had no desire to carry too far a radical revolution. Conciliation was still necessary if he was to succeed to the presidency and retain the support of the army; nor had he ever intended to disappropriate

the leaders of industry, provided they served the interests of the Nazi state. Ernst Röhm was the chief protagonist of the "continuing revolution"; he was also, as head of the SA, greatly distrusted by the army. Hitler tried first to secure Röhm's support for his policies by persuasion and by giving him government office but failed to win him over. Göring and Heinrich Himmler were eager to remove Röhm, but Hitler hesitated until the last moment. Finally, on June 29, 1934, he reached his decision. Röhm and his lieutenant Edmund Heines were executed without trial, together with Gregor Strasser, Schleicher, and others. The army leaders, satisfied at seeing the SA broken up, approved Hitler's actions. When Hindenburg died, on August 2, they, together with Papen, assented to the merging of the chancellorship and the presidency—with which went the supreme command of the armed forces of the *Reich*—and officers and men took an oath of allegiance to Hitler personally. Economic recovery and a reduction in unemployment (coincident with world recovery, but for which Hitler took credit) made the regime more acceptable, and a combination of success and terrorism brought the support of 90 percent of the voters in a plebiscite.

In power, Hitler devoted little attention to the organization and running of the domestic affairs of the Nazi state. Responsible for the broad lines of policy, as well as for the system of terror that upheld the state, he left detailed administration to his subordinates. Each of these exercised arbitrary power in his own sphere, but, by deliberately creating offices and organizations with overlapping authority, Hitler effectively prevented any one of these private empires from ever becoming sufficiently strong to challenge his own absolute authority.

Foreign policy claimed his greater interest. His objectives were laid down in *Mein Kampf,* and Hitler worked toward them with consummate skill. He had early admired the pan-Germanism of the Austrian Georg Ritter von Schönerer, and the reunion of the German peoples was his first ambition. Beyond that, the natural field of expansion lay eastward, in Poland, the Ukraine, and the U.S.S.R.—expansion that would necessarily involve renewal of Germany's historic conflict with the Slav peoples, who would be subordinate in the new order to the Teutonic master race. He regarded Fascist Italy as a natural ally in this crusade against Bolshevism, provided its rivalry with Germany in central Europe could be overcome, and was ready to abandon the Germans of the Tirol to this end. Britain was a possible ally, provided it abandoned its traditional policy of maintaining the balance of power in Europe and limited itself to its interests overseas. France alone in the west was the natural enemy of Germany and must, therefore, be subdued to make expansion eastward possible.

Planning eastward expansion

Before such expansion was possible, it was necessary to remove the restrictions placed on Germany by the Treaty of Versailles. Hitler used all the arts of propaganda to allay the suspicions of the other powers. He posed as the champion of Europe against the scourge of Bolshevism and insisted that he was a man of peace who wished only to remove the inequalities of the Versailles Treaty. Germany withdrew from the Disarmament Conference and from the League of Nations (October 1933), but Hitler hastened to sign a nonaggression treaty with Poland (January 1934). Every repudiation of the treaty was followed by an offer to negotiate a fresh agreement and insistence on the limited nature of Germany's ambitions. Only once did he overreach himself: when the Austrian Nazis, with the connivance of the German embassy, murdered Chancellor Engelbert Dollfuss of Austria and attempted a coup d'etat (July 1934). The attempt failed, and as Mussolini moved troops to the frontier, Hitler disclaimed all responsibility and sacrificed those who had acted with his sanction. In January 1935 a plebiscite in the Saarland returned that territory to Germany, and Hitler took the opportunity to renounce any further claims on France. In March of the same year, he announced the introduction of conscription, and, although this provoked the united opposition of Britain, France, and Italy at the Stresa Conference, his peace propaganda was sufficiently successful to persuade the British to negotiate a naval treaty (June 1935) recognizing Germany's right to rearm. His greatest stroke came in March 1936, when he used the excuse of a pact between France and the Soviet Union to remilitarize the Rhineland—a decision that he took against the advice of his own general staff. Meanwhile, the alliance with Italy, foreseen in *Mein Kampf,* rapidly became a reality as a result of the sanctions imposed by Britain and France against Italy in the Ethiopian war. In October 1936 the Rome–Berlin axis was established; shortly afterward came the Anti-Comintern Pact with Japan; and these two were linked a year later.

By 1937–38 a new stage had been reached. In November 1937 Hitler outlined his plans of future conquest (beginning with Austria and Czechoslovakia) to a secret meeting of his military leaders. He now dispensed with the services of those who were not wholehearted in their acceptance of Nazi dynamism—Hjalmar Schacht, who declared Germany's further rearmament a danger to its economy; Werner von Blomberg and Werner von Fritsch, representatives of the caution of professional soldiers; and Konstantin von Neurath, Hindenburg's appointment at the foreign office.

Austria and Czechoslovakia

In February 1938 Hitler invited the Austrian chancellor, Kurt von Schuschnigg, to Berchtesgaden and forced him to sign an agreement giving the Austrian Nazis virtually a free hand. When Schuschnigg attempted to repudiate the agreement and announced a plebiscite on the question of an *Anschluss* with Germany, Hitler immediately ordered the occupation of Austria by German troops. The enthusiastic reception that Hitler himself received decided him to settle the future of Austria by outright annexation. He returned in triumph to Vienna, the scene of his youthful humiliations and hardships. No resistance was encountered from Britain and France. Hitler had taken special care to secure the support of Italy, and when this was forthcoming proclaimed his undying gratitude to Mussolini.

Having given assurances that the *Anschluss* would not affect Germany's relations with Czechoslovakia, Hitler proceeded at once with his plans against that country. Konrad Henlein, leader of the German minority in Czechoslovakia, was instructed to agitate for impossible demands on the part of the Sudetenland Germans, thereby enabling Hitler to justify the annexation of Czechoslovakia. But the willingness of Britain and France to compel the Czech government to cede the Sudetenland areas to Germany presented Hitler with the choice between substantial gains by peaceful agreement and even greater acquisitions by a spectacular war against Czechoslovakia. Mussolini's intervention appears to have decided him, and he accepted the Munich Agreement on September 30—only to feel resentment immediately afterward at being cheated out of an impressive military conquest.

It was to be expected, therefore, that Hitler would waste no time in provoking an occasion for occupying the whole of Czechoslovakia. This he did by fostering Slovak discontent. On March 16, 1939, from the Hradčany Castle in Prague, he proclaimed the dissolution of the state whose existence he, as an Austrian, had always regarded as unnatural. Immediately afterward the Lithuanian government was forced to cede Memel (Klaipéda), on the northern frontier of East Prussia, to Germany.

Hitler was now ready to advance toward the ultimate objective of *Lebensraum* in the East. Confronted by an uncompromising Poland, guaranteed by Britain and France, he strengthened the alliance with Italy (the "Pact of Steel," May 1939) and negotiated a nonagression pact with the Soviet Union, signed on August 24—just within the deadline set for an attack on Poland before the winter. He still disclaimed any quarrel with Britain, but to no avail, and the invasion of Poland (September 1) was followed two days later by a British and French declaration of war.

Invasion of Poland

In his foreign policy Hitler combined complete opportunism in means and timing with unwavering pursuit of the objectives laid down in *Mein Kampf.* He showed astonishing skill in judging the mood of the democracies and exploiting their weaknesses—in spite of the fact that he had scarcely set foot outside Austria and Germany and spoke no foreign language. Up to this point every move had been successful—even his anxiety over British and French entry into the war was dispelled by the rapid success of the war in Poland. The result was to convince

him more and more of his own infallibility and to induce him to push ahead still faster with his plans for conquest.

World War II. Hitler from the first had assumed direction of the major strategy of the war. When the success of the campaign in Poland failed to lead to the peace negotiations with Britain for which he had hoped, he ordered the army to prepare for an immediate offensive in the west. Bad weather, however, provided the reluctant generals with the opportunity to postpone the western offensive, and this in turn led to two major changes in planning. The first, on the suggestion of Adm. Erich Raeder, commander in chief of the navy, was Hitler's order to occupy Denmark and Norway in April 1940. Hitler took a close personal interest in the operation, and from this time his intervention in the detail of military operations was to grow steadily greater.

The second was the adoption of Gen. Erich von Manstein's plan for an attack through the Ardennes (opened May 10), instead of through the Low Countries. Against his generals' advice, Hitler held back Gen. Heinz Guderian's tanks south of Dunkirk, thus enabling the British to organize the evacuation of their army. But the campaign as a whole was a brilliant success, and Hitler could claim the major credit for its overall planning. On June 10 Mussolini entered the war on the side of Germany, and at the end of June Hitler avenged the Treaty of Versailles by signing an armistice with France on the site of the Armistice of 1918.

The next step was the subjugation of Britain by aerial bombardment, followed by invasion. But in the summer of 1940, long-term preparations were begun for the invasion of the Soviet Union, and, as the expected surrender of Britain still failed to materialize, the eastern campaign quickly came to dominate Hitler's conception of the grand strategy of the war to the exclusion of everything else. The Soviet Union had occupied eastern Poland and Bessarabia, and Hitler sought to counter any further moves by forcing the governments of Hungary and Romania to accept an agreement that he dictated and by urging the abandonment of Mussolini's plans for the invasion of Greece. Mussolini, however, piqued at being kept in ignorance of Hitler's intentions, invaded Greece; and the lack of success of the Italian armies made it necessary for German forces to come to their aid in the Balkans and North Africa. Hitler's plans were further disrupted by a coup d'etat in Yugoslavia in March 1941, overthrowing the government that had made an agreement with Germany. Regarding this as an insult to Germany and himself, Hitler immediately ordered his armies to subdue Yugoslavia. The campaigns in the Mediterranean theatre, although successful, remained subordinate to the eastern offensive, with which Hitler was so preoccupied that he lost the opportunism and flexibility that he had shown in political affairs. Even when Raeder and Erwin Rommel urged Hitler to destroy the whole British Middle East position by a final blow at Suez, he would spare no more forces from Operation "Barbarossa"—the planned invasion of the Soviet Union.

The attack against the U.S.S.R. was launched on June 22, 1941, with Hitler so confident of success that he refused to provide winter clothing and equipment for his troops. The German Army advanced swiftly into the Soviet Union, but failed to destroy its Russian opponent. Hitler became completely overbearing toward his generals. He disagreed with them about the object of the main attack and he wasted time and strength by failing to concentrate on a single objective and by frequently reversing his own decisions. In December 1941 an unexpected Russian counterattack made it clear that Hitler's hopes of a single campaign would not be realized.

War with the United States

The next day came the Japanese attack on Pearl Harbor. Hitler precipitately declared war on the United States—although the pact with Japan was purely defensive and he had not been informed of the Japanese intentions. Misled by an essentially central European view of world politics, he apparently took no account of the force that a mobilized United States could bring to bear in Europe.

Hitler's conduct throughout 1942 was marked by further errors of judgment—he paid insufficient attention to the Mediterranean and the Atlantic at a time when a relatively small additional effort in those theatres might have been decisive. In the Soviet Union his continued unreadiness to concentrate on a single objective probably forfeited the opportunity to capture Stalingrad while it was still lightly defended.

Meanwhile, he directed Himmler to prepare the ground for the "new order" in Europe. The concentration camps were expanded, and there were added to them extermination camps such as Auschwitz and Mauthausen, as well as mobile extermination squads. The Jews of Germany, Poland, and the Soviet Union were most numerous among the victims; in German-occupied Europe between 4,500,000 and 5,500,000 had been killed by the end of the war as the only solution in Hitler's view of the Jewish "problem." The sufferings of other races were only less when measured in numbers killed. Such barbarism was indiscriminate, even where, as in the Ukraine, Hitler might have encouraged nationalist feelings to his own advantage.

At the end of 1942, defeat at el-Alamein and at Stalingrad brought the turning point in the war, and Hitler's character and way of life began to change. Hitherto, the success that he had imagined had been largely realized, but to preserve the world of fantasy from defeat and failure he isolated himself more and more from reality. Directing operations from his headquarters in the east, he refused to visit bombed cities or to read reports of setbacks; those close to him, especially Martin Bormann, his secretary, took care that only pleasing information reached him; and he became increasingly dependent on his physician, Theodor Morell, and the injections that he supplied. Even so, he had not yet lost the power to react vigorously in the face of misfortune. After the arrest of Mussolini in July 1943 and the Italian armistice, he not only directed the occupation of all important positions held by the Italian army but ordered the kidnapping of Mussolini, with the intention that he should head a new Fascist government. On the eastern front, however, the refusal to withdraw led only to greater losses without any possibility of holding up the Soviet advance. Inevitably, relations with his army commanders grew increasingly strained, the more so with the growing importance given to the SS divisions, directly responsible to Hitler. Meanwhile, the failure of the U-boat campaign and the bombing of Germany made more evident how reduced were the chances of victory.

Assassination attempts

All these factors made more desperate the few soldiers and civilians who were ready to remove Hitler and negotiate a peace. Several attempts were planned in 1943–44; the most nearly successful was made on July 20, 1944, when Col. Claus von Stauffenberg exploded a bomb at a conference at Hitler's headquarters in East Prussia. But Hitler escaped with superficial injuries, and, with few exceptions, those implicated in the plot were executed. The reduction of the army's independence was now made complete, and National Socialist political officers appointed to all military headquarters.

Thereafter, Hitler was increasingly ill and fatigued; but he did not relax or lose control over the Nazi Party or the army, and he continued to exercise an almost hypnotic power over his close subordinates, none of whom was able to wield any independent authority. In December 1944 he moved his headquarters to the west to direct an offensive in the Ardennes for which the last reserves of manpower were mobilized. When it failed, his hopes for victory became ever more visionary, based on the use of new weapons or on the breakup of the grand alliance, especially after the death of Roosevelt. Far from trying to save what could be rescued from defeat, he ordered mass material destruction and condemned his armies to death by refusing to allow them to surrender.

From January 1945 Hitler never left the chancellery in Berlin or its bunker, abandoning a plan to lead a final resistance in the south as the Soviet forces closed in on Berlin. In a state of extreme nervous exhaustion, prematurely senile if not insane, he at last accepted the inevitability of defeat and thereupon prepared to take his own life, leaving to its fate the country over which he had taken absolute command. Before this, two further acts remained. Around the midnight of April 28–29 he married Eva Braun. Immediately afterward he dictated his political

testament, justifying his career and appointing Karl Dönitz as head of the state and Josef Goebbels as chancellor.

Hitler's death

On April 30 he said farewell to Goebbels, Martin Bormann, and the few others remaining, then retired to his suite and either shot or poisoned himself. His wife took poison. In accordance with his instructions, their bodies were burned.

Hitler's success must be attributed to the susceptibility of postwar Germany to his own unique talents as a political leader. His rise to power was not inevitable, and any change in a complex conjunction of circumstances might have relegated him to the obscurity and failure of his youth; yet there was no one who equalled his ability to exploit and shape events to his own ends. The power that he wielded was unprecedented, both in its scope and in the technical resources at its command; but he made no permanent contribution, moral or material, to mankind. His originality and distinctiveness lay in his methods rather than his ideas and purpose, which were shared in whole or in part by millions of people, in Germany and elsewhere. By the time he was defeated, he had broken down the whole structure of the world in which he lived and left behind a Germany and a Europe that have remained divided ever since.

(A.B./W.F.Kn.)

BIBLIOGRAPHY

Biographical studies: ALAN BULLOCK, *Hitler: A Study in Tyranny,* rev. ed. (1964); JOACHIM C. FEST, *Hitler* (1974; originally published in German, 1973), a biography by a journalist; ROBERT G.L. WAITE, *The Psychopathic God* (1977), a psychohistorical study; FRANZ JETZINGER, *Hitler's Youth* (1958, reprinted 1976; originally published in German, 1956); BRADLEY F. SMITH, *Adolf Hitler: His Family, Childhood and Youth* (1967); J. SYDNEY JONES, *Hitler in Vienna: 1907–1913* (1982); KONRAD HEIDEN, *Der Fuehrer: Hitler's Rise to Power* (1944, reissued 1969; originally published in German, 1936), which covers the period up to 1934; H.R. TREVOR-ROPER, *The Last Days of Hitler,* 5th ed. (1978). RICHARD F. HAMILTON, *Who Voted for Hitler?* (1982), an analysis of his political support.

Hitler's own words: Hitler's speeches have been collected and edited by MAX DOMARUS, *Reden und Proklamationen, 1932–1945,* 2 vol. (1962–63, reissued 1973); a selection of his speeches in English for the years 1922–39 has been edited by NORMAN H. BAYNES, *The Speeches of Adolf Hitler, April 1922–August 1939,* 2 vol. (1942, reissued 1981); *Secret Conversations, 1941–44* (Eng. trans. 1953, reissued 1976; British title, *Table Talk, 1941–44,* 1953); *Mein Kampf,* 2 vol. (1925–27; Eng. trans. by RALPH MANHEIM, 1969).

Reminiscences of Hitler: ERNST F.S. HANFSTAENGL, *Hitler: The Missing Years* (U.S. title, *Unheard Witness,* 1957), covering the years 1922–34; ALBERT SPEER, *Inside the Third Reich* (1970; originally published in German, 1969).

The History of the Holy Roman Empire

Holy Roman Empire (Latin, Sacrum Romanum Imperium; German, Heiliges Römisches Reich) is the designation commonly given to the varying complex of lands ruled over first by Frankish and then by German kings for 10 centuries from Charlemagne's coronation in 800 until the renunciation of the imperial title in 1806. (For histories of the territories governed at various times by the empire, see FRANCE; GERMANY; ITALY.)

This article is divided into the following sections:

Nature of the empire 629
Origins of the empire and sources of imperial ideas 629
  Religious origins 629
  Political origins 630
Coronation of Charlemagne as emperor 630
Charlemagne's successors 631
  The Carolingian Empire 631
  The Ottonian Empire 631
Empire and papacy 632
  The Investiture Controversy 632
  The Hohenstaufen emperors 633
The empire after Frederick II 633
  The empire in modern times 633
  The end of the empire 634
Bibliography 634

NATURE OF THE EMPIRE

The precise term Sacrum Romanum Imperium dates only from 1254, though the term Holy Empire reaches back to 1157, and the term Roman Empire was used from 1034 to denote the lands under Conrad II's rule. The term Roman emperor is older, dating from Otto II (died 983). This title, however, was not used by Otto II's predecessors, from Charlemagne (or Charles I) to Otto I, who simply employed the phrase *imperator augustus* ("august emperor") without any territorial adjunct. The first title that Charlemagne is known to have used, immediately after his coronation in 800, is "Charles, most serene Augustus, crowned by God, great and pacific emperor, governing the Roman empire." This clumsy formula, however, was soon discarded.

These questions about terms reveal some of the problems involved in the nature and early history of the empire. It can be regarded as a political institution, or approached from the point of view of political theory, or treated in the context of the history of Christendom as the secular counterpart of a world religion. The history of the empire is also not to be confused or identified with the history of its constituent kingdoms, Germany and Italy, though clearly they are interrelated. The constituent territories retained their identity; the emperors, in addition to the imperial crown, also wore the crowns of their kingdoms. Finally, whereas none of the earlier emperors from Otto I had assumed the imperial title until actually crowned by the pope in Rome, after Charles V none was emperor in this sense, though all laid claim to the imperial dignity as if they had been duly crowned as well as elected. Despite these anomalies and others, the empire, at least in the Middle Ages, was by common assent, along with the papacy, the most important institution of western Europe.

Historical theories justifying the empire

Theologians, lawyers, popes, ecclesiastics, rulers, rebels like Arnold of Brescia and Cola di Rienzo, literary figures like Dante and Petrarch, and the practical men, members of the high nobility, on whom the emperors relied for support, all saw the empire in a different light and had their own ideas of its origin, function, and justification. Among these heterogeneous and often incompatible views, three may be said to predominate: (1) the papal theory, according to which the empire was the secular arm of the church, set up by the papacy for its own purposes and therefore answerable to the pope and, in the last resort, to be disposed of by him; (2) the imperial, or Frankish, theory, which placed greater emphasis on conquest and hegemony as the source of the emperor's power and authority and according to which he was responsible directly to God; and (3) the popular, or Roman, theory (the "people" at this stage being synonymous with the nobility and in this instance with the Roman nobility), according to which the empire, following the tradition of Roman law, was a delegation of powers by the Roman people. Of the three theories the last was the least important; it was evidently directed against the pope, whose constitutive role it implicitly denied, but it was also a specifically Italian reaction against the predominance in practice of Frankish and German elements.

It is also important to distinguish between the universalist and localist conceptions of the empire, which have been the source of considerable controversy among historians. According to the former, the empire was a universal monarchy, a "commonwealth of the whole world, whose sublime unity transcended every minor distinction"; and the emperor "was entitled to the obedience of Christendom." According to the latter, the emperor had no ambition for universal dominion; his policy was limited in the same way as that of every other ruler, and when he made more far-reaching claims his object was normally to ward off the attacks either of the pope or of the Byzantine emperor. According to this view, also, the origin of the empire is to be explained by specific local circumstances rather than by far-flung theories.

ORIGINS OF THE EMPIRE AND SOURCES OF IMPERIAL IDEAS

There was no inherent reason why, after the fall of the Roman Empire in the West in 476 and the establishment there of Germanic kingdoms, there should ever again have been an empire, still less a Roman empire, in western Europe. The reason this took place is to be sought (1) in certain local events in Rome in the years and months immediately preceding Charlemagne's coronation in 800, and (2) in certain long-standing tendencies that made this particular solution of a difficult situation thinkable. These long-standing tendencies are to be regarded as preconditions rather than causes of the coronation; they do not account for it, but without them it is difficult to imagine how it could have taken place.

The first is the persistence, despite the fact that it was no longer a political reality west of Italy, of the idea of the universal and eternal Roman Empire. The importance of this tradition may easily be exaggerated. After the establishment of the Germanic kingdoms, kings such as Clovis in Gaul and Theodoric the Great in Italy were glad to accept Roman titles from the Byzantine emperor, who sought to maintain the formal unity of the Roman Empire by treating them as his vicars and lieutenants; but this was a short-lived expedient. By the 8th century any sense of belonging to the empire, by then confined to eastern Europe, had disappeared in the West.

The Christian Empire

Religious origins. Far more effective in the minds of the barbarian peoples of the West was the idea of the *Imperium Christianum* or "Christian Empire," which took shape after the conversion of Constantine the Great and the reconciliation between Christianity and the Roman Empire. Not only did the Christian Church become a state church, including in its liturgy prayers for the empire and the emperor, but it also brought the Roman Empire into the framework of Christian eschatology, as the last of the world monarchies whose end would mark the inception of the kingdom of God. Through Christian iconography and through the liturgy the church's view of the empire as a vehicle of God's will, for the Christianization of the world,

became prevalent. It was expressed with peculiar force in the letters of Charlemagne's adviser Alcuin.

Apart from the persistence of the idea of a Christian Roman Empire, a third precondition for the establishment of an empire in the West was the existence of a candidate of sufficient power and standing in the person of the Frankish king. The Frankish kingdom expanded until it comprised most of western Europe, and it acquired the Lombard kingdom in northern Italy in 774. The importance of the ties forged by Charlemagne's immediate predecessors with the papacy are obvious. Though it is scarcely true that Charlemagne's accession to the empire was simply a consequence of this expansion, his outstanding position was evidently a precondition of his elevation to the imperial throne.

Political origins. When one turns from the preconditions to the causes of the events of 800 and from the realm of ideas to that of political facts, one enters another world. It is the world of the Byzantine (or Eastern) Empire, of which the pope was a subject, confirmed in his bishopric like other bishops by the Byzantine emperor. By the 8th century, however, the imperial position in Italy, centred in the Exarchate of Ravenna, was crumbling. Confronted by the power of Islām, the empire concentrated on the problems of the East and was unable to defend its Italian lands from the Lombards, who had entered Italy in 568. Furthermore, the Byzantine emperors of the Isaurian dynasty, of whom the first was Leo III (717–741), were estranged from the papacy by the Iconoclastic Controversy. Hitherto fear of the Lombards had kept the popes faithful to the Byzantine emperor and to the exarch in Ravenna, but the situation changed. When the Lombards renewed their advance southward, taking Ravenna in 751 and driving out the exarch, the pope appealed in vain to the Byzantine emperor. He then, in an epoch-making turn to the West, sought assistance from the Frankish king, Pepin III the Short, who invaded Italy and reduced the Lombard king Aistulf to submission (754). Returning in 756, Pepin bestowed on the papacy the territories belonging to the exarchate. Thus was born both the temporal power of the papacy and the close alliance between the papacy and the Frankish monarchy.

Alliance of papacy and Frankish kings

Particularly significant of the anomalous position that had arisen in Italy was the action of the pope in conferring on the Frankish king the title of patrician, which could legally be conferred only by the emperor; in the form the pope gave it (*patricius Romanorum*), the title was meant to authorize its possessor to defend and support the Holy See against its foes. Even so, the papacy was reluctant to break its constitutional links with the imperial system, and this was particularly so under Pope Adrian I (772–795), after Charlemagne had in 774 defeated the last independent Lombard king, Desiderius, and taken his kingdom. As king of Lombardy the defender might become as dangerous to the pope as those against whom he had provided defense.

The events briefly related show, first, the growing estrangement between East and West; second, the difficult position of the pope between the Byzantine emperor and the Frankish king; third, the tenuousness of the imperial hold over Rome and central Italy. After the expulsion of the exarch in 751, the pope himself, in de facto possession, probably hoped to become the emperor's successor in the West. Whether the document itself belongs to the 750s or not, this was the sense of the forged Donation of Constantine, in which the emperor Constantine I is represented as having conferred on Pope Sylvester I the imperial palace of the Lateran, the imperial insignia, and "all the provinces, places and cities of Italy and the western regions." Such ambitions, however, proved illusory when Charlemagne became king of the Lombards. When it was even rumoured that Charles intended to depose Adrian and set up a Frankish pope in his place, it would have been folly for the pope to think of formal severance from the existing empire. Nevertheless, the empire's shadowy titles were unlikely to be respected long in Italy unless the Byzantine emperor took action. From 781 papal documents were no longer being dated by the emperor's regnal year, and the pope was minting his own coins.

CORONATION OF CHARLEMAGNE AS EMPEROR

By comparison with Adrian, Pope Leo III (795–816) was a man of inferior calibre. Where Adrian had tried to maintain independence by balancing the Byzantine emperor against the Frankish king, Leo from the first showed subservience to the latter. Both in Constantinople and in Rome the situation was unstable. In Constantinople, after troubles reaching back to 790, the empress Irene had her son Constantine VI blinded and deposed in 797 and took his place, the first woman to rule the empire in her own right. Her constitutional position was thus doubtful; Alcuin in the West, in 799, regarded the imperial throne as empty. Meanwhile, in Rome the hostile nobility exploited the opportunity to attack Leo, who in 799 fled across the Alps to his protector, Charlemagne, at Paderborn. Though unfavourably impressed by the Pope, Charlemagne was persuaded by Alcuin to send him back to Rome with a commission, which adjudged the complaints against him false and arrested and deported his accusers. The situation, however, was still uncertain. In view of the plight of both pope and Byzantine emperor, "the whole salvation of the church of Christ" rested (so Alcuin wrote) in Charlemagne's hands, and in the autumn of 800 he set out for Rome "to restore the state of the church which was greatly disturbed." On December 23 Leo solemnly purged himself of the charges against him. Two days later, on December 25, a large gathering assembled in St. Peter's, where the Pope was to consecrate Charlemagne's son as king. Suddenly, as Charlemagne rose from prayer, Leo placed a crown on his head and, while the assembled Romans acclaimed him as "Augustus and emperor," the Pope abased himself before Charlemagne, "adoring" him "after the manner of the emperors of old."

Crowning of Charlemagne

It seems clear that this coronation was the work of the papacy, not of the Frankish king, who is said to have been surprised and angry at it. The immediate beneficiary of the coronation was the pope, whose position henceforth was secure. Charlemagne was left to face its momentous consequences and, particularly, to secure that recognition from Constantinople without which his title was legally invalid. This, according to the chronicler Theophanes, he sought to do by offering marriage to the empress Irene, hoping thus "to reunite east and west." If so, a revolution in Constantinople and the deposition of Irene in 802 brought the plan to nothing. In any case, the coronation of Charlemagne was an extralegal, indeed an illegal and revolutionary, proceeding. The pope had no right to make him emperor. Nor did the coronation create a new western by the side of the existing eastern empire. A usurper in the eyes of the Byzantines, Charlemagne had not the least prospect of succeeding to the throne of the Caesars. The only imperial territories on which he laid hands were the duchy of Rome and the former exarchate. Otherwise he remained, as before, king of the Franks and of the Lombards. In view of the fact that in 806 he made arrangements to divide his territories among his three sons, one may doubt whether Charlemagne's empire would have survived had not the two elder sons died before him, leaving the undivided inheritance in 814 to the third son, Louis I the Pious.

Although the immediate context of the imperial coronation of 800 was limited, it had wider connotations. In the first place, the separation between East and West had become an accomplished fact in the political sphere; for, though the intention in 800 was not to divide the empire, this was the practical outcome. In 812, after unsuccessful war and wearisome negotiation, the Byzantine emperor Michael I recognized Charlemagne's imperial title. It was still a personal title, and Charlemagne was recognized merely as emperor, not as emperor of the Romans; in other words, the emperor in Constantinople maintained his claim to be the only true successor to the Roman Caesars. Furthermore, the recognition was grudgingly given, and later, when Byzantium was stronger and the Carolingians weaker, Michael's successors refused to extend it automatically to Charlemagne's successors. Thus the second consequence of the act of 800 was a rivalry with Constantinople, which remained an important factor in imperial history at least until 1204. In the third

place, Charlemagne's coronation involved him and his successors ever more deeply in the ecumenical pretensions of the papacy.

The relationship between the papacy and the Frankish rulers, close for nearly 50 years before 800, was intensified when the Roman see became the first metropolitan church of Charlemagne's dominions. Religious emperors and their ecclesiastical advisers would henceforward see as the main function attaching to their imperial dignity the promotion of Christian unity. Furthermore, the fact that the pope had crowned Charlemagne emperor—rightfully or not—could not but impress. It was the pope who had taken the initiative. Had he not, in fact, constituted Charlemagne emperor? In Innocent III's time it was to be argued that Pope Leo III had transferred the empire from the Greeks to the Germans and that his successors could transfer it elsewhere if they so wished. This was a later doctrine; but already to Charlemagne the dangers were evident. Hence when, in 813 after his agreement with Michael I, Charlemagne decided to associate his surviving son, Louis, in the exercise of imperial power, he framed his actions accordingly. The ceremony took place not in Rome but in the imperial chapel at Aachen; the pope was not present; the constitutive act was the acclamation of the gathered Frankish nobility; and Louis either received the diadem from his father or took it with his own hands from the altar. The contrast with the Roman ceremony of 800 was deliberate. Henceforward the conflict between the two contrary views or theories of the empire—the papal and the Frankish—was to be a dominant theme.

Assertion of imperial independence from Rome

CHARLEMAGNE'S SUCCESSORS

The Carolingian Empire. Louis I the Pious (814–840) was a man in every way different from his father. For him the word empire was to be the unifying idea holding together his various dominions, and accordingly he abandoned his separate royal titles. This was the underlying notion of the *Ordinatio imperii* of 817; by this, Louis made his eldest son, Lothair I, emperor with him, while the younger sons, Pepin and Louis the German, received the subordinate kingdoms of Aquitaine and Bavaria. Louis I's ideas, however, aroused the conservative opposition of the Frankish nobility, and soon the Frankish lands were involved in civil war. Furthermore, internal dissension helped the papacy, which increased its influence by favouring one party or the other. Already in 816 Pope Stephen IV (or V) had persuaded Louis to receive unction at his hands at Reims and also to be recrowned, thus repairing the defect of 813; and though Lothair I had been raised to the imperial dignity in 817, as Louis had been in 813, without papal intervention, he also saw fit to strengthen his title by being crowned by Paschal I in Rome in 823. Thus the empire, instead of becoming an appendage of the Frankish kingship, was drawn back into the papal sphere. Lothair's son Louis II was crowned by the pope in 850, to be sole emperor from Lothair's death in 855 to his own in 875; his uncle Charles II the Bald was emperor from 875 to 877; then Charles III the Fat was crowned emperor in 881. But the imperial title, without power to support it, was a mere name, and with the dislocation of the Carolingian realm, culminating in the deposition of Charles the Fat by the East Frankish magnates in 887, its decline was rapid.

Papal crowning of Louis I

From 888 France, Germany, and Italy were separate states, with the kingdom of Burgundy and Lorraine (Lotharingia) as debatable lands. Who, in these circumstances, was to be emperor? The nominee of the pope, himself a puppet of Italian aristocratic factions? Or Charlemagne's rightful heir, whoever he might be? Or simply the strongest king of western Europe? For the moment, the first solution prevailed, and there followed a series of emperors chosen from the Italian nobility, starting with Guy of Spoleto (891–894) and his son Lambert (894–898), in concurrence with the East Frankish king Arnulf, who was also crowned emperor in 896 (and died 899). But such men were of little use to the papacy; and the papacy, in the depths to which it had sunk, was little use to them. Louis III, crowned in 901, was deposed in 905 by Berengar of Friuli, who was himself crowned by the pope in 915, but on Berengar's death in 924 the powerful Roman family of the Crescentii, determined to keep authority in its own hands, stepped in and suppressed the imperial title. Thus the empire created in 800 disappeared, ineffective and unmourned.

The Ottonian Empire. While in Italy the empire was in process of dissolution, north of the Alps imperial ideas persisted, but not in the form associated with the papacy. The Frankish tradition of empire looked back to Charlemagne; amid the divisions of the dying 9th century it reflected the notion of Frankish unity. It was also seen as the expression of power and hegemony and of rule—as Charlemagne had ruled—over a number of different





countries. Thus the West Frankish king Charles II the Bald had been proclaimed emperor and Augustus when he invaded Lorraine in 869—six years before he was crowned emperor in Rome—because he was now ruler of two kingdoms. The first Saxon king of Germany, Henry I, was acclaimed emperor (the chronicler Widukind relates) because of his victory over the Hungarians in 933, as was his son, Otto I, after the Battle of Lechfeld in 955. This is a Frankish idea of empire that has nothing to do with Rome. Indeed, the successor-states that are called France and Germany remained Frankish—the West Frankish and the East Frankish kingdoms, though the latter was ruled by a Saxon. As such, it was their ambition to succeed to the Frankish inheritance, and it was inevitable, in view of the anarchic state of Italy, that the ruler of the one or the other, when strong enough, would cross the Alps. In fact, it was the East Frankish kingdom that recovered first from the 9th-century anarchy, not least because its kings knew how to use the Frankish church for their ends.

In this Otto I, who succeeded to the East Frankish throne in 936, was the true heir of Charlemagne; he made churchmen his ministers and established missionary bishoprics on the Elbe River to spread Christianity among the Wends. In the west the Carolingian Louis IV of France was his protégé, and only Otto's support kept the young king of Burgundy, Conrad, on his throne. In 951 Otto descended on Italy, married the late king's widow, and forced the new king, Berengar of Ivrea, to hold his kingdom as a vassal of the German crown. Ten years later, summoned by the pope when Berengar was on the point of taking control in Rome, Otto returned to Italy. He was crowned emperor by Pope John XII on Candlemas day 962.

Once again, as in 800, it was the pope who took the initiative; his need for protection was the immediate cause of the coronation. But Otto's empire was more limited in scope than Charlemagne's, its pretensions less universal. In the intervening period the import of the title *imperator augustus* had shrunk, so that by Otto's time it scarcely extended beyond protection or "advocacy" of the Roman church. Likewise the extent of the empire, territorially, had decreased. Neither Otto I nor his successors made any claim, as heirs of Charlemagne, to rule over the West Frankish lands. Henceforward what came to be known as the empire meant simply a union of Germany and northern Italy (and after 1032 the kingdom of Burgundy) under a single rule. This was far from the ecumenical pretensions of Rome, and though, partly because of the extent of their territories, partly because of their special connection with the papacy, the Ottonian and Salian emperors enjoyed preeminence among the rulers of western Europe, they certainly did not seek to exercise world dominion. Nor did they seek to challenge the Byzantine emperors in their rule over the empire of the East. Nevertheless the problem of the relations of the two empires necessarily reasserted itself, for the Byzantine Empire had made rapid recovery and was preparing, under Nicephorus II Phocas and Basil II Bulgaroctonus, to embark on expansion and reconquest. Fears of Byzantine interference in Roman politics and conflicting claims in southern Italy soon brought the two powers into rivalry, and the conflict proved to be the spur for a major development—the romanization of the western empire. Whereas Otto I had laid no claim to the Roman title, Otto II (co-emperor 967; sole emperor 973–983), to bolster his claims against Basil II, proclaimed himself Roman emperor. From this it was only a short step to describing the empire itself as the Roman Empire, and this change occurred in Conrad II's reign 1024–39; crowned 1027). It was reinforced, from 1040, by the introduction of the title "king of the Romans" for the emperor-elect before his coronation, or for the emperor's designated successor. It is noteworthy, however, that these formal changes were the result not of claims to the heritage of ancient Rome or of universal pretensions but simply of rivalry with Byzantium.

"King of the Romans"

In this development the reign of Otto III (983–1002; crowned 996) forms an interlude. The son of a Greek princess, imbued with Byzantine traditions, Otto III drew on Roman, Carolingian, and Christian strands to form a new synthesis combining the heterogeneous elements in the imperial idea. Rome was to be his capital, the pope his lieutenant in the spread of Christian dominion. How much of this would have survived, if Otto had lived longer, remains doubtful. In fact, distance alone made unified imperial rule over Germany and Italy impossible. Otto himself lost Rome shortly before his death. What did survive, as the characteristic feature of the empire from Otto I to Henry III, was the close subordination of pope to emperor. Otto I deposed John XII and Benedict V; Otto III nominated his cousin, Gregory V, and his tutor, Sylvester II; and after neglect of Italian affairs under Henry II (1002–24; crowned 1014) and Conrad II, Henry III (1039–56; crowned 1046) deposed Benedict IX and Sylvester III, and compelled Gregory VI to abdicate. This was not oppression or force but the fulfillment of the emperor's supreme duty—to watch over the welfare of the church. The iconography of the time shows the emperor, not the pope, as the representative of St. Peter and often standing alone as God's vicar on earth. The period from 962 to 1046, in fact, saw the empire at its zenith. No other kingdom of Europe was its match; while the papacy, except when under imperial protection, was the degraded tool of Roman factions. Morally, also, the empire had the support of the greatest reformers of the age—the abbots of Cluny, for example—who looked to the emperor rather than to the pope as the effective head of Latin Christendom.

EMPIRE AND PAPACY

From the middle of the 11th century the situation began to change. One cause was the rapid progress of European economic recovery, which brought shifts of power detrimental to Germany. More immediately important was the revival of the papacy, which the emperors had done so much to further. After Henry III's death in 1056 the initiative passed into papal hands. It was favoured by the long minority—until 1065—of Henry IV (crowned 1084; died 1106), which enabled the papacy to act without fear of intervention from north of the Alps, and by the appearance of allies—particularly the Normans of the Kingdom of Sicily, who for their own purposes supported the papacy against the empire. As they reached maturity the peoples of Europe turned to the pope as leader of Christendom. Even within the imperial frontiers the emperor's power meant more to the Germans than to the inhabitants of Burgundy or of Italy, for whom it betokened subjection to German rule. Furthermore, only Otto III—and he for less than four years—made Rome the seat of empire; all the rest, from Charlemagne onward, concentrated their efforts north of the Alps. In practice, therefore, the empire was a very imperfect realization of the ideal of an *imperium Christianum;* and as soon as it was in a position to vindicate its independence, the papacy found many adherents.

The Investiture Controversy. Under Pope Gregory VII (1073–85) the papal theory of the empire, as formulated in the 9th century, was revived, but on broader and firmer foundations. The result was the conflict, from 1076 until 1122, known as the Investiture Controversy, ostensibly centring on the question of whether lay overlords had the authority to "invest" bishops and abbots within their domains—that is, to appoint them and formally give them the symbols of their office. The real issue, however, was not the investiture of bishoprics and abbacies but the place of the emperor in Christian society and his relations with the papacy. Only the pope, Gregory VII asserted, might use the imperial insignia; he might lawfully depose emperors but should himself be judged by none (these lapidary statements are among the 27 included in the *Dictatus papae* of 1075 and were set down in Gregory's register). Thus the claim to independence turned rapidly into a claim to superiority. In particular, the sacred character of the emperor was challenged, as was his claim to be responsible directly to God. Instead, on the basis of the Donation of Constantine and a papal interpretation of the coronation of 800, it was argued that it was for the pope to convey the imperial dignity and, if he thought fit, to withhold or withdraw it. The Investiture Controversy was brought to a close by compromise in the Concordant of Worms of 1122 between Pope Calixtus II and the emperor Henry V;

but Gregory VII's claims were taken up again by popes Alexander III, Innocent III, Innocent IV, and Boniface VIII, in a series of conflicts that shook the empire to its foundations.

The Hohenstaufen emperors. The challenge thrown out by Gregory VII forced the emperors to seek new foundations for their position. Gregory's great opponent, the emperor Henry IV, had still asserted the traditional rights of his father. His successors in the 12th century, Henry V (1106–25; crowned 1111), Lothair II (1125–37; crowned 1133), Frederick I Barbarossa (1152–90; crowned 1155), and Henry VI (1190–97; crowned 1191), shifted their ground. To counter the arguments of church lawyers they grasped the weapons provided by the revival of Roman law. A new and more exalted conception of the empire was the result. Best known was the addition by Frederick I Barbarossa, in 1157, of the word *sacrum* to the name of the empire, which then became the Sacrum Imperium (Holy Empire) as a counterblast to the Sancta Ecclesia (Holy Church). Equally characteristic was the canonization of Charlemagne by Frederick's antipope Paschal III in 1165. In this way Frederick emphasized continuity with the Frankish past and asserted his rights as Charlemagne's successor. They derived, he argued, not from conferment by the pope or by the Roman people but from Frankish conquest.

"Holy Roman Empire"

Unlike earlier emperors, who had based their position on their special relation with the church, the Hohenstaufen emperors emphasized its secular foundations. Against Pope Innocent III's claims to confer the imperial crown, imperial lawyers asserted that "he who is chosen by the election of the princes alone is the true emperor, even before he has been confirmed by the pope." Nor is it surprising that, confronted with the universal claims of the papacy, the Hohenstaufen emperors asserted rights no less universal. Though in day-to-day politics, in their relations with the kings of France or of England, for example, there is no sign that they were seeking world dominion, nevertheless the new imperialism soon called forth protests from all sides—from England and France, from Denmark and Hungary. "Who," asked John of Salisbury, "appointed the Germans to be judges over the nations?"

Adventures in Italy

Meanwhile, the conflict with the papacy and the desire to restore the territorial basis of imperial power, which the Investiture Controversy had shattered, drew the emperors more and more into Italy, where they encountered the same national reaction. Unable to defeat the Lombard League, a northern Italian urban coalition, Frederick I patched up the Peace of Constance in 1183. His ultimate sovereignty was recognized, but his power in Italy was fatally compromised. After his son, Henry VI, had through marriage inherited the kingdom of southern Italy and Sicily, the power of the Norman kingdom was used to restore the imperial position in Italy. It was a grandiose policy but overstrained. The papacy, fearing that Rome would be engulfed, reacted violently.

Pope Innocent III, profiting from German dissensions after the early death of Henry VI (1197), played upon the German factions (Otto IV, not established as king until 1208, was crowned emperor in 1209). Henry VI's son Frederick II (1212–50; crowned 1220), by the *Privilegium in favorem principum ecclesiasticorum* (1220) and by the *Statutum in favorem principum* (1232), made far-reaching concessions to the German princes in order to ensure their support for his Italian policy, but in vain. In spite of his striking victory at Cortenuova in 1237 Frederick II failed to crush the Lombards and was excommunicated in 1239 and deposed in 1245. His death in 1250 marked the effective end of the medieval empire. In Germany a long interregnum (from 1250 to 1273) brought down the imperial structure. In Italy, to ensure that there could be no restoration, the papacy called in Charles of Anjou, a younger son from the French royal house, who conquered the south and became King Charles I of Naples and Sicily (1266–85). When Rudolf I of Habsburg succeeded as German king in 1273, he was only the head of a federation of princes, while in Italy he abandoned all claims over the centre and south, and he retained only titular rights in Lombard.

THE EMPIRE AFTER FREDERICK II

The uncrowned emperors

It is characteristic of the new situation that Rudolf I of Habsburg, though he made a number of attempts, never formally achieved the imperial dignity. Henceforward the title of emperor, though it continued, usually did not have the sanction of personal crowning by a pope or papal legate. For a century after Frederick II's death the only "true" emperor was Henry VII (king from 1308 to 1313), who was crowned in Rome in 1312 by legates of the Avignon pope. Thereafter until the end of the empire there were in all only four emperors who were duly crowned: Charles IV, crowned by a legate in 1355; Sigismund, by the pope in 1433; Frederick III, in 1452; and Charles V, by the pope but at Bologna, in 1530. If the empire and imperial title continued to exist, it resulted partly from the force of tradition, partly from the exigencies of German politics, and partly from fear of the dangerous conflict of interests that any plan for its abolition would necessarily involve.

The Germans, naturally, were unwilling to surrender hope of regaining something of the empire's former power: both Henry VII and Louis IV (king from 1314 to 1347; his Roman coronation in 1327 was by representatives of the people) sought to revive the Italian policies of the Hohenstaufen. But the balance had swung against them. France was already striving for the imperial position that Napoleon was ultimately to secure, and France determined that the Germans should not recover the imperial prerogatives. Moreover, in Germany itself, civil war had undermined the power of the kingship, and the elective monarchy was effectively controlled by the princes through the college of electors definitely established soon after 1250. French pretensions to leadership in Europe provoked a last tardy revival of imperialist sentiment both in Germany (Alexander of Roes at the end of the 13th century, Engelbert of Admont at the beginning of the 14th) and in Italy (Marsilius of Padua and Dante), but the emperor Charles IV, a sober realist, drew the necessary conclusions. By then the axiom that "the king is emperor in his kingdom" was firmly established; it marked the end of any universalist dream. Charles set out accordingly to make the empire a specifically German institution. By agreement with Pope Clement V, he formally abandoned Italy; he would enter Rome only on the day fixed for his coronation and leave again the same day. This he did on April 5, 1355. Then he turned to the definition of the German constitution, particularly the rights of the electors, in the Golden Bull of 1356. The change was reflected in the final evolution of the empire's title: Sacrum Romanum Imperium Nationis Germanicae (Holy Roman Empire of the German Nation). This title, which appears under Frederick III (king from 1440, emperor from 1452 to 1493), indicates that the emperor's powers were limited to his German lands. In 1508 Frederick's successor Maximilian I, unable to go to Rome, assumed with papal consent the style "elected emperor" or "chosen emperor" (Latin *imperator electus;* German *erwählter Kaiser*).

The empire in modern times. The history of the empire after the promulgation of the Golden Bull may be treated briefly, because from that time it is essentially a part of German history. It is true that memories of an imperial past continued to have an influence on German thinking and that in the Habsburg lands there was a sense of belonging to a multinational empire. A few emperors—Sigismund in the 15th century, Charles V in the 16th—may even have thought to recover part of the old imperial prerogative. It was also possible to make something of the empire's leadership of Christendom against the Turks. But institutionally the role of the empire was almost continuously whittled away. After the failure of the project of imperial reform sponsored in 1495 by the elector of Mainz, Berthold of Henneberg, the hope vanished of endowing the empire with permanent institutions effective beyond the limits of the different principalities. The Reformation entrenched the princes firmly in their rights and accentuated their autonomy.

When Charles V, opening the Diet of Worms in 1521, declared that "the empire from of old had not many masters, but one, and it is our intention to be that one,"

he was shutting his eyes to the realities. The extent of his dominions was imposing, but they were a weak dynastic agglomeration; and though Charles championed the Roman Catholic Church against the Reformation, his empire was neither in spirit nor in fact a revival of the medieval empire. When he accepted the Peace of Augsburg in 1555 and abdicated in 1556, the change that was begun with the accession of Rudolf I of Habsburg was completed.

Religious divisions of the Reformation

With Germany split into two religious camps, the emperor was little more than the head of a religious faction. Furthermore, after Sigismund's death (1437), with one short intermission for Charles VII from 1742 to 1745, the imperial crown, though in theory elective, was hereditary in the Habsburg dynasty of Austria; and this fact produced a cleavage of interests between emperor and empire.

The end of the empire. From 1556 until its end under Francis II in 1806 the empire meant little more than a loose federation of the different princes of Germany, lay and ecclesiastical, under the presidency of the House of Habsburg. After the Thirty Years' War (1618–48), no emperor again attempted, as Charles V had done, to reestablish a strengthened central authority; and the Peace of Westphalia in 1648 marked the empire's final organization on federal lines. Yet, even at the end, the empire had loyal adherents, particularly among the small knights and noblemen of western Germany, who regarded it as their safeguard against princely absolutism; and its role was not so entirely negative as is sometimes thought. Its loose structure still suited to some degree the cosmopolitan spirit of the 18th century. But with the French Revolution, and the intensified nationalism that followed, it became an anachronism.

French rivalry

As far back as the end of the 13th century, French kings had been scheming to annex the title as well as to absorb the outlying territories of the empire. With Napoleon this ambition came within reach. Posing as the new Charlemagne ("because, like Charlemagne, I unite the crown of France to that of the Lombards, and my empire marches with the east"), he resolved in 1806 to oust Francis II from his title and to make the Holy Roman Empire a part of the Napoleonic "new order." He was anticipated, however, by Francis II, who in 1804 had assumed the title "hereditary emperor of Austria" and who, resolving that no other should wear the crown that he was powerless to defend, resigned the old imperial dignity on August 6, 1806.

So perished the Holy Roman Empire. The extent and character of its influence will always be a matter for debate, but it left a deep imprint on Europe. Nor did it cease to be influential after its extinction. The debate about the medieval empire was an ideological background to the creation of the Second Reich, or German Empire, in 1871, and even Hitler's Third Reich drew sustenance from memories, often thwarted and perverted, of Charlemagne and Otto the Great and Frederick II. (G.Bh.)

BIBLIOGRAPHY. The classical account in English is that of JAMES BRYCE, *The Holy Roman Empire* (1864; new ed. 1904, corrected 1906, reprinted 1978). See also GEOFFREY BARRACLOUGH, *The Crucible of Europe: The Ninth and Tenth Centuries in European History* (1976), and *The Medieval Empire: Idea and Reality* (1950, reprinted 1964), somewhat critical of Bryce; and ROBERT FOLZ, *The Concept of Empire in Western Europe from the Fifth to the Fourteenth Century* (1969, reprinted 1980; originally published in French, 1953), with bibliography. On the components of the imperial idea see FEDOR SCHNEIDER, *Rom und Romgedancke im Mittelalter* (1925, reprinted 1959); H.X. ARQUILLIÈRE, *L'Augustinisme politique,* 2nd rev. ed. (1955, reprinted 1972); and CARL ERDMANN, *Forschungen zur politischen Ideenwelt des Frühmittelalters* (1951). On relations of the eastern and western empires: WERNER OHNSORGE, *Das Zweikaiserproblem im früheren Mittelalter* (1947); FRANZ DÖLGER, *Byzanz und die europäische Staatenwelt* (1953, reprinted 1976); GEORGE OSTROGORSKY, *History of the Byzantine State,* rev. ed. (1969, reprinted 1980; originally published in German, 1940). For the Carolingian empire see KARL HELDMANN, *Das Kaisertum Karls des Grossen* (1928); LOUIS HALPHEN, *Charlemagne and the Carolingian Empire* (1977; originally published in French, 1947). For the period of the Investiture see ELIE VOOSEN, *Papauté et pouvoir civil à l'époque de Grégoire VII* (1927); GERD TELLENBACH, *Church, State, and Christian Society at the Time of the Investiture Contest* (1940). On the Hohenstaufen empire see WALTHER HOLTZMANN, *Das mittelalterliche Imperium und die werdenden Nationе* (1953); HANS JOACHIM KIRFEL, *Weltherrschaftsidee und Bündnispolitik* (1959). For the empire after the interregnum, see FRIEDRICH BOCK, *Reichsidee und Nationalstaaten* (1943). The classical account of the imperial constitution after the Thirty Years' War is SAMUEL PUFENDORF, *Die Verfassung des deutschen Reiches* (1976; originally published in Latin, 1667).

(G.Bh./Ed.)

The Homeric Epics

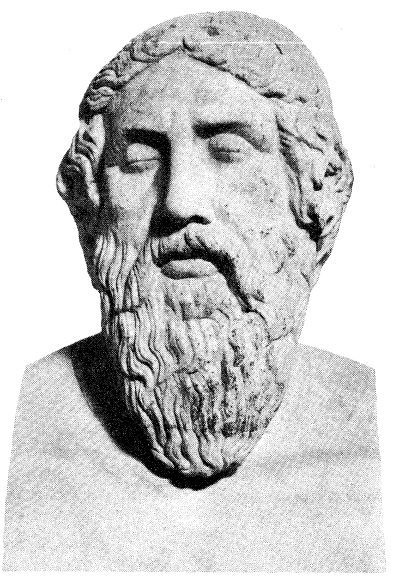
The two great epic poems of ancient Greece, the *Iliad* and the *Odyssey*, have always been attributed to a shadowy figure by the name of Homer. Little is known of him beyond the fact that his was the name attached in antiquity by the Greeks themselves to the two great poems. That there was an epic poet called Homer and that he played the primary part in shaping the *Iliad* and the *Odyssey*—so much may be said to be probable. If this assumption is accepted, then Homer must assuredly be one of the greatest of the world's literary artists.

Homer's influence

He is also one of the most influential authors in the widest sense, for the two epics provided the basis of Greek education and culture throughout the classical age and formed the backbone of humane education down to the time of the Roman Empire and the spread of Christianity. Indirectly through the medium of Virgil's *Aeneid* (which was loosely molded after the patterns of the *Iliad* and the *Odyssey*), directly through their revival under Byzantine culture from the late 8th century AD onward, and subsequently through their passage into Italy with the Greek scholars who fled westward from the Ottomans, the Homeric epics had a profound impact on the Renaissance culture of Italy. Since then the proliferation of translations has helped to make them the most important poems of the classical European tradition.

It was probably through their impact on classical Greek culture itself that the *Iliad* and the *Odyssey* most subtly affected Western standards and ideas. The Greeks regarded the great epics as something more than works of literature; they knew much of them by heart, and they valued them not only as a symbol of Hellenic unity and heroism but also as an ancient source of moral and even practical instruction.

Early references. Implicit references to Homer and quotations from the poems date to the middle of the 7th century BC. Archilochus, Alcman, Tyrtaeus, and Callinus in the 7th century and Sappho and others in the early 6th adapted Homeric phraseology and metre to their own purposes and rhythms. At the same time scenes from the epics became popular in works of art. The pseudo-Homeric "Hymn to Apollo of Delos," probably of late 7th-century composition, claimed to be the work of "a blind man who dwells in rugged Chios," a reference to a tradition about Homer himself. The idea that Homer had descendants known as "Homeridae," and that they had taken over the preservation and propagation of his poetry, goes back at least to the early 6th century BC. Indeed, it was not long before a kind of Homeric scholarship began: Theagenes of Rhegium in southern Italy toward the end of the same century wrote the first of many allegorizing interpretations. By the 5th century biographical fictions were well under way; the Pre-Socratic philosopher Heracleitus of Ephesus made use of a trivial legend of Homer's death—that it was caused by chagrin at not being able to solve some boys' riddle about catching lice—and the concept of a contest of quotations between Homer and Hesiod (after Homer the most ancient of Greek poets) may have been initiated in the Sophistic tradition. The historian Herodotus assigned the formulation of Greek theology to Homer and Hesiod and claimed that they could have lived no more than 400 years before his own time, the 5th century BC. This should be contrasted with the superficial assumption, popular in many circles throughout antiquity, that Homer must have lived not much later than the Trojan War about which he sang.



Restored bust thought to represent Homer, copied from a Greek original, *c.* 450 BC. In the Sala delle Muse, the Vatican.
By courtesy of the Vatican Museum

Personal history

The general belief that Homer was a native of Ionia (the central part of the western seaboard of Asia Minor) seems a reasonable conjecture for the poems themselves are in predominantly Ionic dialect. Although Smyrna and Chios early began competing for the honour (the poet Pindar, early in the 5th century BC, associated Homer with both), and others joined in, no authenticated local memory survived anywhere of someone who, oral poet or not, must have been remarkable in his time. The absence of hard facts puzzled but did not deter the Greeks; the fictions that had begun even before the 5th century BC were developed in the Alexandrian era in the 3rd and 2nd centuries BC (when false scholarship as well as true abounded) into fantastic pseudobiographies, and these were further refined by derivative scholars under the Roman Empire. The longest to have survived purports to be by Herodotus himself; but it is quite devoid of objective truth.

Modern inferences. Modern scholars agree with the ancient sources only about Homer's general place of activity. The most concrete piece of ancient evidence is that his descendants, the Homeridae, lived on the Ionic island of Chios. Yet an east Aegean environment is suggested for the main author of the *Iliad* by certain local references in the poem; that is, to the peak of Samothrace just appearing over the intervening mass of Imbros when seen from the plain of Troy, to the birds at the mouth of the Cayster near Ephesus, to storms off Icaria and northwest winds from Thrace. East Aegean colouring is fainter in the *Odyssey,* which is set primarily in western Greece; but the poem's vagueness over the position of Ithaca, for example, is not incompatible with the idea of a poet in Ionia elaborating materials derived from the farther side of the Greek world.

Admittedly, there is some doubt over whether the *Iliad* and the *Odyssey* were even composed by the same main author. Such doubts began in antiquity itself and depended mainly on the difference of genre (the *Iliad* being martial and heroic, the *Odyssey* picaresque and often fantastic), but they may be reinforced by subtle differences of vocabulary even apart from those imposed by different subjects. Aristotle's conception of the *Odyssey* as a work of Homer's old age is not impossible; but if the *Iliad* is the earlier of the two (as seems likely from its simpler structure and the greater frequency of relatively late linguistic forms in the *Odyssey*), then the *Odyssey* could have been created after its image, and as a conscious supplement, once the example of monumental composition had been given. In any case the similarities of the two poems are partly due

to the coherence of the heroic poetical tradition that lay behind both.

Date of the epics

The internal evidence of the poems is of some use in determining when Homer lived. Certain elements of the poetic language, which was an artificial amalgam never exactly reproduced in speech, indicate that the epics were not only post-Mycenaean in composition but also substantially later than the foundation of the first Ionian settlements in Asia Minor of about 1000 BC. The running together of adjacent short vowels and the disappearance of the semivowel digamma (a letter formerly existing in the Greek alphabet) are the most significant indications of this. At the other end of the time scale the development in the poems of a true definite article, for instance, represents an earlier phase than is exemplified in the poetry of the middle and late 7th century. Both stylistically and metrically, the Homeric poems appear to be earlier than the Hesiodic poems, which many scholars place not long after 700 BC. A different and perhaps more precise criterion is provided by datable objects and practices mentioned in the poems. Nothing, except for one or two probably Athenian additions, seems from this standpoint to be later than around 700; on the other hand, the role assigned in the *Odyssey* to the Phoenicians as traders, together with one or two other phenomena, suggests a date of composition—for the relevant contexts at least—of after 900. A few passages in the *Iliad* may imply a new form of fighting in close formation, dependent on the development of special armour for foot soldiers (hoplites) after about 750, and references to the Gorgon mask as a decorative motif point in the same direction. It is true that the poems contain many traditional and archaic elements, and their language and material background are a compound of different constituents originating at different dates. It seems, nonetheless, plausible to conclude that the period of composition of the large-scale epics (as distinct from their much shorter predecessors) was the 9th or 8th century, with several features pointing more clearly to the 8th. The *Odyssey* may belong near the end of this century, the *Iliad* closer to its middle. It may be no coincidence that cults of Homeric heroes tended to spring up toward the end of the 8th century, and that scenes from the epic begin to appear on pots at just about the same time.

Homer as an oral poet. But even if his name is known and his date and region can be inferred, Homer remains primarily a projection of the great poems themselves. Their qualities are significant of his taste and his view of the world, but they also reveal something more specific about his technique and the kind of poet he was. It has been one of the most important discoveries of Homeric scholarship, associated particularly with the name of an American scholar, Milman Parry, that the Homeric tradition was an oral one—that this was a kind of poetry made and passed down by word of mouth and without the intervention of writing. Indeed Homer's own term for a poet is *aoidos,* "singer." The *Odyssey* describes two such poets in some detail: Phemius, the court singer in the palace of Odysseus in Ithaca, and Demodocus, who lived in the town of the semi-mythical Phaeacians and sang both for the nobles in Alcinous' palace and for the assembled public at the games held for Odysseus. On this occasion he sings of the illicit love affair of Ares and Aphrodite in a version that lasts for exactly 100 Homeric verses. This and the other songs assigned to these singers—for example, that of the Trojan Horse, summarized in the *Odyssey*—suggest that ordinary *aoidoi* in the heroic tradition worked with relatively short poems that could be given completely on a single occasion. That is what one would expect, and it is confirmed by the habits of singers and audiences at other periods and in other parts of the world (the tradition of the poet-singers of Muslim Serbia has provided the most fruitful comparison so far). Whatever the favoured occasion for heroic song—whether the aristocratic feast, the religious festival, or popular gatherings in tavern or marketplace—a natural limitation on the length of a poem is imposed by the audience's available time and interest as well as by the singer's own physique and the scope of his repertoire. Such relatively short songs must have provided the backbone of the tradition inherited by Homer, and his portraits of Demodocus and Phemius are likely to be accurate in this respect. What Homer himself seems to have done is to introduce the concept of a quite different style of poetry, in the shape of a monumental poem that required more than a single hour or evening to sing and could achieve new and far more complex effects, in literary and psychological terms, than those attainable in the more anecdotal and episodic songs of his predecessors.

The poet as *aoidos*

Poetic techniques. It can be asked how one can be so confident in classing Homer himself as an oral singer, for if he differed from Phemius or Demodocus in terms of length, he may also have differed radically in his poetic techniques. The very nature of his verse may provide a substantial part of the answer. The style of the poems is "formulaic"; that is, they rely heavily on the use not only of stock epithets and repeated verses or groups of verses—which can also be found to a much lesser extent in a literate imitator like Virgil—but also on a multitude of fixed phrases that are employed time and time again to express a similar idea in a similar part of the verse. The clearest and simplest instance is the so-called noun-epithet formulas. These constitute a veritable system, in which every major god or hero possesses a variety of epithets from which the choice is made solely according to how much of the verse, and which part of it, the singer desires to use up. Odysseus is called divine Odysseus, many-counseled Odysseus, or much-enduring divine Odysseus simply in accordance with the amount of material to be fitted into the remainder of the hexameter (six-foot) verse. A ship is described as black, hollow, or symmetrical not to distinguish this particular ship from others but solely in relation to the qualities and demands of the rhythmical context. The whole noun-epithet system is both extensive and economical—it covers a great variety of subjects with very little exact reduplication or unnecessary overlap. It would seem that so refined and complex a system could not be the invention of a single poet but must have been gradually evolved in a long-standing tradition that needed both the extension and the economy for functional reasons—that depended on these fixed phrase units because of its oral nature, in which memory, practice, and a kind of improvising replace the deliberate, self-correcting, word-by-word progress of the pen-and-paper composer. Admittedly, the rest of Homer's vocabulary is not as markedly formulaic as its noun-epithet aspect (or, another popular example, as its expressions for beginning and ending a speech). Many expressions, many portions of sentences are individually invented for the occasion, or at least so it seems. Even so, there is a strongly formulaic and ready-made component in the artificial language that was used by Homer, including its less conspicuous aspects such as the arrangement of particles, conjunctions, and pronouns.

It looks, therefore, as though Homer must have trained as an ordinary *aoidos,* who began (like most of the present-day Yugoslav *guslari*) by building up a repertoire of normal-length songs acquired from already established singers. The greatest heroic adventures of the past must already have been prominent in any repertoire, especially the Panhellenic adventures of the Seven Against Thebes, the Argonauts, and the Achaean attack on Troy. Some aspects of the Trojan War might already have been expanded into songs of unusual length, though one that was still manageable on a single occasion. Yet the process was presumably carried much further in the making of the monumental *Iliad,* consisting of more than 16,000 verses, which would take four or five long evenings, and perhaps more, to perform. This breakthrough into the monumental, which made exceptional and almost unreasonable demands of audiences, presupposes a singer of quite exceptional capacity and reputation—one who could impose the new and admittedly difficult form upon his listeners by the sheer unfamiliar genius of his song. The 8th century BC was in other respects, too, an era of cultural innovation, not least in the direction of monumentality, and huge temples (like the early temple of Hera in Samos) and colossal funerary vases (like the mixing bowls and amphorae in the so-called Geometric style from the Dipylon cemetery in Athens) may have found a literary analogue in the idea of a vast poetical treatment of the Trojan War. But in an

The monumental poem

important sense Homer was building upon a tendency of all known oral heroic poetry toward elaboration and expansion. The singer does not acquire a song from another singer by simple memorization. He adjusts what he hears to his existing store of phrases, typical scenes, and themes, and he tends to replace what is unfamiliar to him with something he already knows, or to expand it by adding familiar material that it happens to lack. Every singer in a living oral tradition tends to develop what he acquires. There is an element of improvisation, as well as of memory, in his appropriation of fresh material; and judging by the practice of singers studied from the middle of the 19th century onward in Russia, Serbia, Cyprus, and Crete the inclination to adjust, elaborate, and improve comes naturally to all oral poets.

Typical elements of an oral poem

Cumulative poetic structure. Homer must have decided to elaborate his materials not only in quality but also in length and complexity. All oral poetry is cumulative in essence; the verse is built up by adding phrase upon phrase, the individual description by adding verse upon verse. The whole plot of a song consists of the progressive accumulation of minor motifs and major themes, from simple ideas (such as "the hero sets off on a journey" or "addresses his enemy") through typical scenes (such as assemblies of men or gods) to developed but standardized thematic complexes (such as episodes of recognition or reconciliation). Homer seems to have carried this cumulative tendency into new regions of poetry and narrative; in this as in other respects (for example, in his poetical language) he was applying his own individual vision to the fertile raw material of an extensive and well-known tradition.

Establishing Homer's contributions

The result is much more complex than with an ordinary traditional poem. Understanding the origin and essential qualities of the *Iliad* or the *Odyssey* entails trying to sort out not only the separate components of the pre-Homeric tradition but also Homer's own probable contributions, whether distinguishable by their dependence on the monumental idea or by their apparent novelty vis-à-vis the tradition as a whole or by other means. Dialectal and linguistic components must be identified as far as possible—survivals of the Mycenaean language, for example, or words used exclusively in the Aeolian cities of the west coast of Asia Minor, or Athenian dialect forms introduced into the poems after the time of Homer; so must specific references to armour, clothing, houses, burial customs, political geography, and so on, that are likely to be assignable to the Late Bronze Age, the Early Iron Age, or the period of Homer's own activity—at the very least to be taken as relatively early or late within the whole range of the poetic tradition down to Homer. These are the tasks of modern Homeric scholarship. Yet such different forms and ideas in Homer are not conveniently separated into distinct sections of the text, which can therefore be assigned to early or late phases of composition. On the contrary, they may coexist in a single (artificial) linguistic form or a single descriptive phrase. Any member of the tradition, not least Homer himself, may, moreover, have chosen to archaize on one occasion, to innovate on another. One result is that the epics are dubious authorities for the assessment of historical events like the attack on Troy or the status of workers, just as they are ambiguous sources for early Greek grammar or theology. Another is that they are not bound to a single worldview or period or mode of perception; rather, they unite judgments and experiences never seen together in "real" life into a whole that is literary but nevertheless revealing of the underlying structure of human existence.

Stabilizing the text. An important and difficult question, which affects the accuracy of modern Homeric texts, is that of the date when the epics became "fixed"—which means given authoritative written form, since oral transmission is always to some extent fluid. An alphabetic writing system reached Greece in the 9th or early 8th century BC; before that was a gap of 200 or 300 years, following the collapse of Mycenaean culture and the disappearance of Linear B writing (with each sign generally representing a syllable), during which Greece seems to have been nonliterate. During that interval, certainly, much of the epic tradition was formed. The earliest alphabetic inscriptions to have survived, a few of them containing brief scraps of hexameter verse, date from around 730 BC. Therefore, if Homer created the *Iliad* at some time after 750 BC, he could conceivably have used writing to help him. Some scholars think that he did. Others believe that he may have remained nonliterate (since literacy is not normally associated with oral creativity) but dictated the poem to a literate assistant. Still others believe that the poems may have been preserved orally and not too inaccurately at least until the middle years of the following, the 7th, century, when "literature" in the strict sense appeared in the poetry of Archilochus. There are objections to all three theories, but this much can be generally agreed: that the use of writing was in any case ancillary, that Homer behaved in important ways like a traditional oral poet. Some scholars are convinced that certain of the more subtle effects and cross-references of Homer's poetry would be impossible without the ability to consult a written text. That is doubtful; certainly the capacities even of ordinary oral poets in this direction are constantly surprising to habitual literates.

At least it may be accepted that partial texts of the epics were probably being used by the Homeridae and by professional reciters known as rhapsodes (who were no longer creative and had abandoned the use of the lyre) by the latter part of the 7th century BC. The first complete version may well have been that established as a standard for rhapsodic competitions at the great quadrennial festival at Athens, the Panathenaea, at some time during the 6th century BC. Even that did not permanently fix the text, and from then on the history of the epics was one of periodical distortion followed by progressively more effective acts of stabilization. The widespread dissemination of the poems consequent upon the growth of the Athenian book trade in the 5th century and the proliferation of libraries after the 4th was followed by the critical work of the Alexandrian scholar Aristarchus of Samothrace in the 2nd century BC, and much later by the propagation of accurate minuscule texts (notably the famous manuscript known as Venetus A of the *Iliad*), incorporating the best results of Greco-Roman scholarship, in the Byzantine world of the Middle Ages. Rare portions of either poem may have been added after, but not long after, the main act of composition; the night expedition that results in the capture of the Trojan spy Dolon and that fills the 10th book of the *Iliad,* some of the underworld scenes in the 11th book of the *Odyssey,* and much of the ending of the *Odyssey* after line 296 of the 23rd book (regarded by Aristarchus as its original conclusion) are the most probable candidates on the grounds of structure, language, and style.

Even apart from the possibilities of medium-scale elaboration, the *Iliad* and the *Odyssey* exemplify certain of the minor inconsistencies of all oral poetry, and occasionally the composer's amalgamation of traditional material into a large-scale structure shows through. Yet the overriding impression is one of powerful unity.

The story of the *Iliad*

The Iliad. The *Iliad* is not merely a distillation of the whole protracted war against Troy but simultaneously an exploration of the heroic ideal in all its self-contradictoriness—its insane and grasping pride, its magnificent but animal strength, its ultimate if obtuse humanity. The poem is, in truth, the story of the wrath of Achilles, the greatest warrior on the Greek side, that is announced in its very first words; yet for thousands of verses on end Achilles is an unseen presence as he broods among his Myrmidons, waiting for Zeus's promise to be fulfilled—the promise that the Trojans will set fire to the Achaean ships and force King Agamemnon to beg him to return to the fight. Much of the poetry between the first book, in which the quarrel flares up, and the 16th, in which Achilles makes the crucial concession of allowing his friend Patroclus to fight on his behalf, consists of long scenes of battle, in which individual encounters alternate with mass movements of the opposing armies. The battle poetry is based on typical and frequently recurring elements and motifs, but it is also subtly varied by highly individualized episodes and set pieces: the catalog of troop contingents, the formal duels between Paris and Menelaus and Ajax and Hector, Helen's identifying of the Achaean princes, Agamemnon

inspecting his troops, the triumph of Diomedes, Hector's famous meeting back in Troy with his wife Andromache, the building of the Achaean wall, the unsuccessful embassy to Achilles, the night expedition, Hera's seduction of Zeus and Poseidon's subsequent invigoration of the Achaeans. Patroclus' death two-thirds of the way through the poem brings Achilles back into the fight, although not before the recovery of Patroclus' body, the making of new divine armour for Achilles, and his formal reconciliation with Agamemnon. In book 22 he kills the deluded Hector; next he restores his heroic status by means of the funeral games for Patroclus; and in the concluding book Achilles is compelled by the gods to restore civilized values and his own magnanimity by surrendering Hector's body to King Priam.

The Odyssey. The *Odyssey* tends to be blander in expression and sometimes more diffuse in the progress of its action, but it presents an even more complex and harmonious structure than the *Iliad.* The main elements are the situation in Ithaca, where Penelope, Odysseus' wife, and their young son, Telemachus, are powerless before her arrogant suitors as they despair of Odysseus' return from the siege of Troy; Telemachus' secret journey to the Peloponnese for news of his father, and his encounters there with Nestor, Menelaus, and Helen; Odysseus' dangerous passage, opposed by the sea-god Poseidon himself, from Calypso's island to that of the Phaeacians, and his narrative there (from book 9 to book 12) of his fantastic adventures after leaving Troy, including his escape from the cave of the Cyclops, Polyphemus; his arrival back in Ithaca, solitary and by night, at the poem's halfway point, followed by his meeting with his protector-goddess Athena, his elaborate disguises, his self-revelation to the faithful swineherd Eumaeus and then to Telemachus, their complicated plan for disposing of the suitors, and its gory fulfillment. Finally comes the recognition by his faithful Penelope, his recounting to her of his adventures, his meeting with his aged father, Laertes, and the restitution, with Athena's help, of stability in his island kingdom of Ithaca. (See also GREEK LITERATURE: *Epic narrative.*)

The story of the *Odyssey*

Homer's influence seems to have been strongest in some of the most conspicuous formal components of the poems. The participation of the gods can both dignify human events and make them seem trivial—or tragic; it must for long have been part of the heroic tradition, but the frequency and the richness of the divine assemblies in the *Iliad,* or the peculiarly personal and ambivalent relationship between Odysseus and Athena in the *Odyssey,* probably reflect the taste and capacity of the main composer. The many-sidedness of battle, the equivocal realism of death in a hundred forms, must have been developed among Homer's predecessors but can never before have been deployed with such massive and complex effect. In the extended similes the strain of heroic action is relieved by the illuminating intrusion of a quite different and often peaceful contemporary world, in images developed often almost longingly beyond the immediate point of comparison. These similes, in their placing and their detail at least, surely depend on the main composer. And yet, beyond such general intuitions as these, the attempt to isolate his special contributions often becomes self-defeating. The *Iliad* and the *Odyssey* owe their unique status precisely to the creative and therefore unanalyzable confluence of tradition and design, the crystalline fixity of a formulaic style and the mobile spontaneity of a brilliant personal vision. "Homer" implies, above all, this fusion.

The result is an impressive amalgam of literary power and refinement. The *Iliad* and the *Odyssey,* however, owe their preeminence not so much to their antiquity and to their place in Greek culture as a whole but to their timeless success in expressing on a massive scale so much of the triumph and the frustration of human life. Although all literature must be engaged with that to some degree, epic poems are not where one most expects to find it. But these poems rise above the immediate concerns of heroic battle or the struggle against gods and nature or against monstrous forces, and they do so with the help of a poetical language of great simplicity and subtlety, a rugged and surprisingly variable narrative technique, and a nucleus of remarkable tales set around the Trojan War and its aftermath. Their greatest power lies, perhaps, in their dramatic quality because much of each poem consists of conversation and speeches, in which rhetoric is kept firmly under control and the individual characters emerge as they confront each other and the gods with advice, inquiry, request, resignation, and passion. Achilles, Hector, Menelaus, Ajax, Odysseus, and the others acquire a kind of heroic glow that even Greek tragedy later found hard to emulate. That is the result, in part, of the very archaism of these age-old tales, which the special techniques of monumental composition never attempted to conceal; but it also depends on something that overlaps that archaism, namely a sheer mythic quality imparting to these tales something of the universal validity to which all great literature aspires and which Homer achieved consistently and with an apparent ease that must be deceptive.

BIBLIOGRAPHY

Greek text: DAVID B. MONRO and THOMAS W. ALLEN (eds.), *Homeri Opera,* vol. 1–2, *Ilias,* 3rd ed. (1920, reprinted 1978), and vol. 3–4, *Odyssea,* 2nd ed. (1916, reprinted 1975–79).

Translations: The best close modern translations are RICHMOND LATTIMORE (trans.), *The Iliad of Homer* (1951, reprinted 1976), and *The Odyssey of Homer* (1967, reprinted 1975); and WALTER SHEWRING (trans.), *The Odyssey* (1980). A loose but powerful contemporary verse translation is ROBERT FITZGERALD (trans.), *The Iliad* (1974, reissued 1984).

Commentaries: G.S. KIRK (ed.), *The Iliad, a Commentary* (1985–), the first of a projected six-volume series; M.M. WILLCOCK (ed.), *The Iliad of Homer, Books I–XII* (1978), and *The Iliad of Homer, Books XIII–XXIV* (1984); and W.B. STANFORD (ed.), *Ὁμήρου Ὀδυσσεία: The Odyssey of Homer,* 2nd ed., 2 vol. (1958–59, reissued 1973–74). See also the Italian commentaries on the *Odyssey* by S. WEST (Books 1–4), J.B. HAINSWORTH (Books 5–8), A. HEUBECK (Books 9–12), and A. HOEKSTRA (Books 13–16) (1981–84).

Critical studies: W.A. CAMPS, *An Introduction to Homer* (1980); JASPER GRIFFIN, *Homer* (1980), an introduction for the general reader, and *Homer on Life and Death* (1980, reprinted 1983); HOWARD CLARKE, *Homer's Readers: A Historical Introduction to the "Iliad" and the "Odyssey"* (1981); NORMAN AUSTIN, *Archery at the Dark of the Moon: Poetic Problems in Homer's "Odyssey"* (1975); JAMES M. REDFIELD, *Nature and Culture in the "Iliad": The Tragedy of Hector* (1975); M.I. FINLEY, *The World of Odysseus,* 2nd rev. ed. (1977, reissued 1982); G.S. KIRK, *The Songs of Homer* (1962, reprinted 1977), abbreviated as *Homer and the Epic* (1965, reprinted 1974); G.S. KIRK (ed.), *The Language and Background of Homer: Some Recent Studies and Controversies* (1964, reissued 1967); ALBERT B. LORD, *The Singer of Tales* (1960, reissued 1978); PAUL MAZON, *Introduction à l'Iliade* (1943, reprinted 1967); DENYS L. PAGE, *History and the Homeric Iliad* (1959), and *The Homeric Odyssey* (1955, reprinted 1976); MILMAN PARRY, *L'Épithète traditionnelle dans Homère* (1928); ALAN J.B. WACE and FRANK H. STUBBINGS (eds.), *A Companion to Homer* (1962, reprinted 1974); T.B.L. WEBSTER, *From Mycenae to Homer* (1958, reprinted 1977); and W.J. WOODHOUSE, *The Composition of Homer's Odyssey* (1930, reprinted 1969).

Allied studies: JOHN CHADWICK, *The Decipherment of Linear B,* 2nd ed. (1968, reprinted 1970); M.I. FINLEY, *Early Greece: The Bronze and Archaic Ages,* rev. ed. (1981); and LORD WILLIAM TAYLOUR, *The Mycenaeans,* rev. ed. (1983).

(G.S.K.)

Hong Kong

Hong Kong (Wade–Giles romanization Hsiang-kang, Pinyin Xianggang) is a British colony and a major trade and financial centre located to the east of the Chu Chiang (Pearl River) Estuary on the south coast of China. It consists of Hong Kong Island, ceded by China to Britain in 1842, the southern part of the Kowloon Peninsula and Stonecutters Island, ceded in 1860, and the New Territories, which include the mainland area lying largely to the north, together with 230 large and small offshore islands, which were leased from China for 99 years in 1898. With the signing of the Chinese–British joint declaration on Dec. 19, 1984, both countries agreed that the entire territory would be returned to China on July 1, 1997.

Hong Kong has 412 square miles (1,067 square kilometres) of land area, including land reclaimed from the sea, and the area continues to grow as more land is reclaimed. Hong Kong Island and its adjacent islets have an area of only about 35 square miles, while urban Kowloon, which includes the Kowloon Peninsula south of Boundary Street, and Stonecutters Island measure only about six square miles. The New Territories account for the rest of the area, amounting to more than 90 percent of the total. The Victoria urban district located on the barren rocks of the northwestern coast of Hong Kong Island is the place where the British first landed in 1841, and it has since been the centre of administrative and economic activities.

Although Hong Kong developed initially on the basis of its excellent natural harbour and the lucrative China trade, particularly opium dealing, it was the expansion of its territory that provided labour and other resources necessary for sustained commercial growth. The community remains limited in space and natural resources, and it faces persistent problems of overcrowding, trade fluctuations, and social and political unrest. Nevertheless Hong Kong has emerged strong and prosperous, albeit with a changed role, as an entrepôt, a manufacturing and financial centre, and a vital agent in the trade and modernization of China. In the modern world Hong Kong is well known not only for its trade and commerce but also as a model of urban development in Asia.

This article is divided into the following sections:

Physical and human geography 639
- The land 639
  - Relief
  - Drainage
  - Soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 640
  - Ethnic composition
  - Linguistic composition
  - Religions
  - Demographic trends
- The economy 642
  - Resources
  - Agriculture and fishing
  - Industry
  - Finance
  - Trade
  - Transportation
- Government and social conditions 643
  - Government
  - Education
  - Housing
  - Health and welfare
- Cultural life 644
  - Cultural milieu and the arts
  - Cultural institutions
  - Recreation
  - Press and broadcasting

History 644
- Early settlement 644
- Events before and during World War II 645
- Modern Hong Kong 645

Bibliography 645

Physical and human geography

THE LAND

Relief. In sharp contrast to the low-lying areas of the Chu Chiang delta, but conforming geologically and structurally to the great South China massif, a well-eroded upland region, Hong Kong has rugged relief and marked variations in topography. Structurally the area is an anticline, running northeast–southwest, that was formed toward the latter part of the Jurassic Period. Lava poured into this structure and formed volcanic rocks that were later intruded by an extensive granitic dome. The harbour of Hong Kong was formed by the drowning of the denuded centre of the dome. The surrounding hills on the mainland and on Hong Kong Island are partly capped by volcanic rocks, and steep, scarplike concave slopes lead to the inner harbour.

Geologic characteristics

The area is a partially submerged, dissected upland terrain that rises abruptly to heights of more than 2,950 feet (900 metres); its backbone is made up of a series of ridges, running northeast to southwest, that tie in closely with the structural trend in South China. This trend is clearly observable from the alignment of Lantau Island and the Tolo Channel. From the highest peak, Tai Mo Shan (3,140 feet), the series of ridges extends southwestward to Lantau Island, where the terrain rises to 3,064 feet on Lantau Peak and 2,851 feet on Sunset Peak. Extending southeastward from Tai Mo Shan, the Kowloon Peak attains an altitude of almost 2,000 feet, but there is an abrupt drop to about 650 feet at Devil's Peak. Victoria (Hong Kong) Harbour is well protected by mountains on Hong Kong Island that include Victoria Peak in the west, which rises to 1,818 feet, and Mt. Parker in the east, which reaches a height of about 1,740 feet.

Lowlands of the Hong Kong region, including floodplains, river valleys, and reclaimed land, occupy less than 20 percent of the land. Extensive lowland regions are found only north of Tai Mo Shan, in the Yuen Long and Sheung Shui plains. The urban area that spans the two sides of the harbour, with continued reclamation, takes up only about 10 percent of the level area.

Drainage. Hong Kong lacks a river system of any scope, the only exception being in the north where the Sham Chun River, which forms the boundary between China and Hong Kong, flows into Deep Bay after collecting a number of small tributaries. Most of the streams are small, and they generally run perpendicular to the northeast–southwest trend of the terrain. The construction of reservoirs and their catchment systems has reduced the amount of fresh water available downstream.

Soils. In general, Hong Kong's soils are acidic and are of low fertility. An exception is the alluvial soils, which are found mainly in the Deep Bay area, where the sediment-laden waters of the Chu Chiang meet saline waters at high tide and slow down to deposit their sediments to form mud flats. Paddy rice farming and, more recently, intensive vegetable cultivation have modified the alluvial soils. Elsewhere, hill soils, classified as red-yellow podzolic

and krasnozem, abound. Under forest, these hill soils have a well-developed profile, with rich topsoil, but when they are exposed, as is mostly the case, they tend to be thin and lacking in nutrients. Under tropical conditions, sheet and gully erosion is extensive and drastic.

Climate. Hong Kong lies at the northern fringe of the tropical zone. Its seasonal changes are well marked, however, with hot, humid summers and cool, dry winters. The climate is largely controlled by the pressure systems over the adjacent great Asian landmass and ocean surface. Thus monsoonal winds blow from the northeast in winter as a result of the cooling of the landmass and the development of a large thermal anticyclone over Inner Mongolia. Southeast winds develop in summer when the North Pacific Ocean heats up more slowly through solar radiation and becomes a high-pressure area. The mean January and July temperatures are about 60° F (16° C) and 84° F (29° C), respectively. The lowest recorded temperature was 32° F (0° C) in January 1893, and the highest was 97° F (36° C) in August 1900. Frost occasionally occurs on hilltops in winter. The average annual rainfall amounts to about 85 inches (2,224 millimetres), more than one-half of which falls during the summer months of June, July, and August; only about 10 percent falls from November to March. Tropical cyclones, or typhoons, generally occur between June and October, and of the 20 to 30 typhoons formed over the western North Pacific and South China Sea each year, an average of five or six may affect Hong Kong. The torrential downpours and strong winds that frequently accompany the typhoons sometimes devastate life and property in Hong Kong.

Typhoons

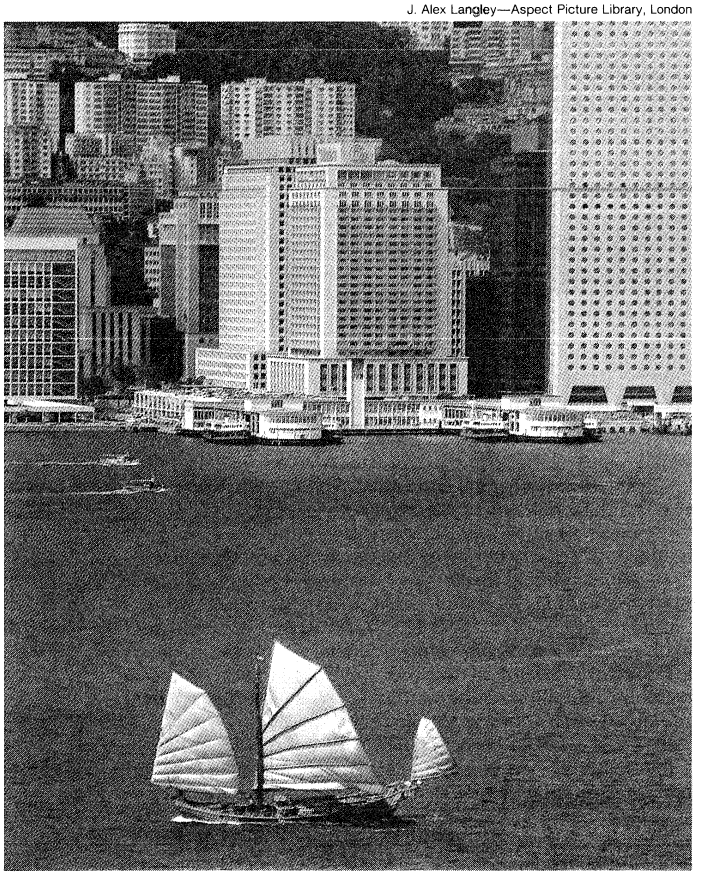
Plant and animal life. Hong Kong is noted for the lushness and great diversity of its plant life. The transitional climate between humid subtropical and warm temperate maritime excludes the most sensitive humid tropical genera due to the cool, dry winter conditions, but many tropical as well as temperate zone families are represented. Most of the land, except for the heavily eroded badlands, is under tropical herbaceous growth, including mangrove and other swamp cover. The most common forest genus today is *Pinus,* represented by the native South China red pine and the slash pine, introduced from Australia. Some of the oldest areas of woodland are in the *feng-shui* wood, or "sacred groves," found in many New Territories villages. These woods consist essentially of native forest trees, some of which are of potential value to the villagers. Centuries of cutting and burning have, however, destroyed much of Hong Kong's original vegetation, leaving only about 13 percent of the land forested. A large portion of Hong Kong's present-day forest cover owes its origin to postwar afforestation programs, which have restored some of the stands of pine, eucalyptus, banyan, casuarina, and palm trees.

Vegetation loss and afforestation

Hong Kong's animal life consists of a mixture of mammals adapted to the subtropical environment. Among the few arboreal mammals are two species of nonnative monkeys that flourish in forests of the New Territories, the rhesus macaque and the longtailed macaque. Tigers are reputed to have once roamed the area, but they are no longer in evidence. The largest remaining carnivores are rare and include the South China red fox, the Chinese leopard cat, the seven-banded civet, and the masked palm civet. Some rat and mouse species typically inhabit scrubland and grassland areas. Birdlife is abundant, and there are numerous snakes, lizards, and frogs.

Settlement patterns. The predominantly urban settlements of Hong Kong are typically distributed linearly, following the irregular coastline and transportation routes. The principal urban areas are established on Hong Kong Island and Kowloon Peninsula, where about three-fourths of the total population lives. There, most of the population is concentrated around Victoria Harbour, living on the limited flatland that is being continuously extended by reclamation. Many major streets, especially those on the northern shore of Hong Kong Island, as well as the entire industrial district of Kwun Tong and the eastern portion of Tsim Sha Tsui at the tip of the peninsula, have been built on reclaimed land.

J. Alex Langley—Aspect Picture Library, London



Junk in Hong Kong (Victoria) Harbour with the Central District of Victoria in the background.

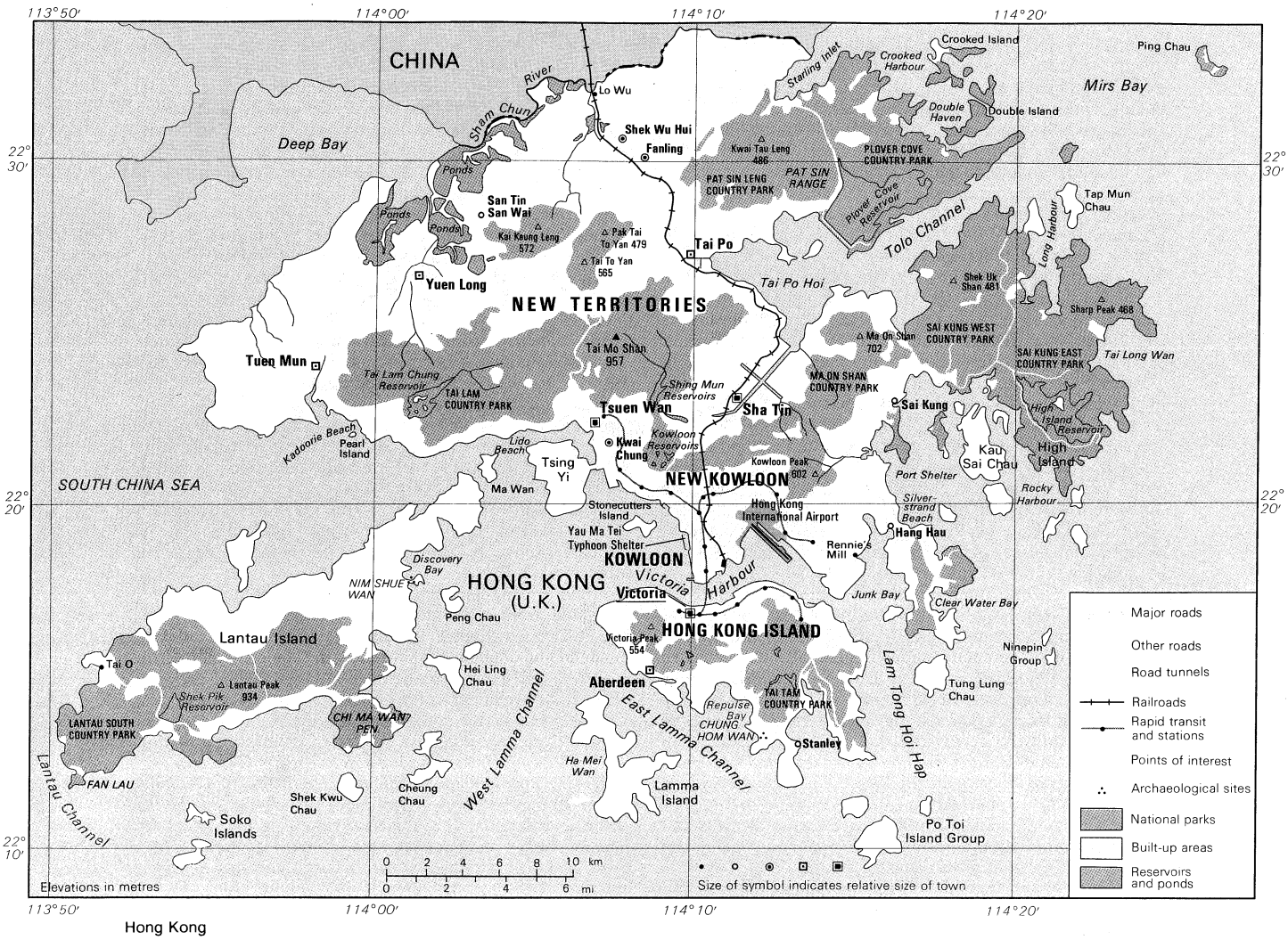
In the New Territories north of the Kowloon hills, rural settlements vary from hamlets to small towns. Most of the villages, compactly built and often walled, follow the alignment of the river systems in the low-lying but fertile alluvial floodplains or the major route corridors. Villages of the Cantonese people are located mainly in the flat alluvial regions, whereas villages of the Hakka people usually are found in narrow valleys or on foothills. The *feng-shui* grove and pond are characteristic of both the Cantonese and Hakka villages; the grove is generally planted on the upslope, or back side, of a village for shade and protection, and the pond is for fish-farming. As a result of population pressure, a number of new towns have sprung up in the New Territories. Among these are Tsuen Wan, Tuen Mun (Castle Peak before 1973), and Sha Tin, which were established in the 1960s and are designed to have populations of between 500,000 and 1,000,000 each. Others, including Tai Po, Fanling, Yuen Long, and, more recently, Junk Bay, were designated as new towns in the 1970s, with population goals ranging from 226,000 to 380,000. Thus, the New Territories, where only 13 percent of the population resided in 1961, accounted for more than one-fourth of the total by the 1980s, and about three-fourths of the New Territories population was concentrated in the new towns.

The new towns

True to its original character as a fishing port, Hong Kong has a sizable, though rapidly dwindling, marine settlement. The "boat people," or Tanka as they are locally known, are essentially fisherfolk living on junks and boats, as their ancestors did for centuries before them. They inhabit fishing towns such as Aberdeen, Shau Kei Wan, and Cheung Chau and typhoon shelters in the harbour areas. With the advance of urbanization and the decline of fishing activity, increasing numbers of them are working ashore. By the early 1980s, this marine population accounted for only about 1 percent of the population, and the decreasing trend was expected to continue.

THE PEOPLE

Ethnic composition. The great majority of the population is Chinese by place of origin, the non-Chinese mak-



Hong Kong

MAP INDEX

Districts

Hong Kong Island . . . . . 22·15n 114·11e
Kowloon . . . . . 22·19n 114·11e
New Kowloon . . . . 22·20n 114·10e
New Territories. . . . 22·24n 114·10e

Cities and towns

Aberdeen . . . . . 22·15n 114·09e
Fanling . . . . . 22·30n 114·08e
Hang Hau . . . . . 22·19n 114·16e
Kwai Chung . . . . . 22·22n 114·08e
Lo Wu. . . . . 22·32n 114·07e
Rennie's Mill . . . . . 22·18n 114·15e
Sai Kung . . . . . 22·23n 114·15e
San Tin San Wai . . . 22·29n 114·03e
Sha Tin . . . . . 22·23n 114·11e
Shek Wu Hui . . . . . 22·30n 114·07e
Stanley . . . . . 22·13n 114·12e
Tai O . . . . . 22·15n 113·51e
Tai Po . . . . . 22·27n 114·10e
Tsuen Wan . . . . . 22·22n 114·07e
Tuen Mun . . . . . 22·24n 113·58e
Victoria . . . . . 22·17n 114·09e
Yuen Long . . . . . 22·26n 114·02e

Physical features and points of interest

Castle Peak Monastery . . . . . 22·23n 113·57e
Cheung Chau, *island*. . . . . 22·12n 114·01e
Chi Ma Wan Peninsula . . . . . 22·14n 114·00e
Chinese University of Hong Kong . . . 22·26n 114·12e
Chung Hom Wan, *archaeological site* . . . . . 22·13n 114·12e
Clear Water Bay . . . 22·17n 114·18e
Crooked Harbour. . . 22·33n 114·16e
Crooked Island . . . . 22·33n 114·18e
Deep Bay . . . . . 22·27n 113·55e
Discovery Bay . . . . 22·18n 114·01e
Double Haven, *anchorage* . . . . . 22·31n 114·18e
Double Island. . . . . 22·31n 114·18e
East Lamma Channel. . . . . 22·14n 114·09e
Fan Lau, *peninsula* . . . . . 22·12n 113·51e
Ha Mei Wan, *bay*. . . 22·12n 114·07e
Hei Ling Chau, *island*. . . . . 22·15n 114·02e
High Island . . . . . 22·22n 114·21e
High Island Reservoir . . . . . 22·23n 114·21e
Hong Kong International Airport. . . . . 22·20n 114·12e
Junk Bay. . . . . 22·17n 114·15e
Kadoorie Beach . . . 22·23n 113·59e
Kai Keung Leng, *peak* . . . . . 22·28n 114·05e
Kau Sai Chau, *island*. . . . . 22·22n 114·19e
Kowloon Peak . . . . 22·21n 114·13e
Kowloon Reservoirs . . . . . 22·21n 114·09e
Kwai Tau Leng, *peak* . . . . . 22·31n 114·12e
Lamma Island. . . . . 22·12n 114·07e
Lam Tong Hoi Hap, *channel* . . . . . 22·15n 114·15e
Lantau Channel. . . . 22·10n 113·50e
Lantau Island . . . . . 22·15n 113·59e
Lantau Peak. . . . . 22·15n 113·55e
Lantau South Country Park. . . . 22·13n 113·51e
Lido Beach . . . . . 22·22n 114·05e
Long Harbour. . . . . 22·27n 114·20e
Mai Po Marshes, *wildlife sanctuary* . . . . . 22·30n 114·03e
Ma On Shan, *peak* . . . . . 22·25n 114·15e
Ma On Shan Country Park. . . . 22·24n 114·15e
Ma Wan, *island*. . . . 22·21n 114·03e
Mirs Bay . . . . . 22·33n 114·24e
Nim Shue Wan, *archaeological site* . . . . . 22·18n 114·01e
Ninepin Group, *islands*. . . . . 22·16n 114·21e
Pak Tai To Yan, *peak* . . . . . 22·28n 114·08e
Pat Sin Leng Country Park. . . . 22·29n 114·12e
Pat Sin Range. . . . . 22·29n 114·13e
Pearl Island . . . . . 22·22n 113·59e
Peng Chau, *island* . . 22·15n 114·02e
Ping Chau, *island* . . 22·33n 114·26e
Plover Cove Country Park. . . . 22·30n 114·16e
Plover Cove Reservoir . . . . . 22·28n 114·15e
Po Lin Monastery . . 22·15n 113·55e
Port Shelter, *bay* . . . 22·21n 114·17e
Po Toi Island Group . . . . . 22·11n 114·16e
Repulse Bay, *beach* . . . . . 22·14n 114·12e
Rocky Harbour . . . . 22·20n 114·19e
Sai Kung East Country Park. . . . 22·24n 114·21e
Sai Kung West Country Park. . . . 22·25n 114·18e
Sham Chun River . . 22·30n 114·02e
Sharp Peak . . . . . 22·26n 114·22e
Shek Kwu Chau, *island*. . . . . 22·12n 113·59e
Shek Pik Reservoir . . . . . 22·14n 113·54e
Shek Uk Shan, *mountain* . . . . . 22·27n 114·18e
Shing Mun Reservoirs . . . . . 22·23n 114·09e
Silverstrand Beach . . . . . 22·20n 114·16e
Soko Islands . . . . . 22·10n 113·54e
South China Sea . . . 22·13n 114·20e
Starling Inlet. . . . . 22·32n 114·14e
Stonecutters Island . . . . . 22·19n 114·08e
Tai Lam Chung Reservoir . . . . . 22·23n 114·01e
Tai Lam Country Park . . . . . 22·24n 114·03e
Tai Long Wan, *bay* . . . . . 22·24n 114·24e
Tai Mo Shan, *mountain* . . . . . 22·25n 114·07e
Tai Po Hoi, *harbour* . . . . . 22·26n 114·12e
Tai Tam Country Park. . . . 22·14n 114·13e
Tai To Yan, *peak* . . . 22·17n 114·08e
Tap Mun Chau, *island*. . . . . 22·29n 114·22e
Tolo Channel . . . . . 22·28n 114·17e
Trappist Monastery . . . . . 22·17n 114·00e
Tsing Yi, *island* . . . . 22·21n 114·05e
Tung Lung Chau, *island*. . . . . 22·15n 114·17e
University of Hong Kong . . . . . 22·17n 114·08e
Victoria Harbour . . . 22·17n 114·10e
Victoria Peak . . . . . 22·17n 114·08e
West Lamma Channel. . . . . 22·13n 114·04e
Yau Ma Tei Typhoon Shelter . . . . . 22·19n 114·10e
Yim Tso Ha Egret Sanctuary. . . . . 22·32n 114·12e

ing up only about 2 percent of the total. Non-Chinese groups are split fairly evenly between non-Asians and Asians. British, Americans, Australians, Canadians, and New Zealanders are among the non-Asians, while the Asian minority groups include Japanese, Indians, Pakistanis, and Singaporeans. An overwhelming majority of the Chinese are from Kwangtung Province and from Hong Kong itself, while less than 10 percent come from other parts of China, notably Fukien, Shanghai, Chekiang, and Kiangsu, and from Taiwan.

Linguistic composition. Although English and Chinese are both official languages, English is used for all legal and governmental matters. Chinese, especially Cantonese in the spoken form, is the common language, however, and is almost universally understood. A variety of dialects and other languages are used among the ethnic minorities. Apart from Cantonese, common dialects, such as Siyi, Chaochow, Hakka, Hoklo, and Tanka, are popularly used within separate native communities of the Kwangtung and Hong Kong Chinese. Groups from other parts of China are also likely to use their own native dialects, and *p'u-t'ung-hua* (Pinyin *putonghua*), or Mandarin, is gaining popularity in Hong Kong. Similarly, the non-Chinese are likely to use their own native languages among themselves.

Religions. The religious persuasions of the people of Hong Kong are as various as their languages and dialects. Among the Chinese, followers of Buddhism and Taoism by far outnumber other groups, and the numerous Buddhist and Taoist temples and monasteries, some centuries old, play an important role in the daily life of the average Chinese. Although each temple is generally dedicated to one or two deities, it is not unusual to find images of a number of other gods or goddesses inside. For a fishing and trading port, the most significant deities are those associated with the ocean and the weather, such as T'ien Hau, the goddess of heaven and protector of seafarers, who is honoured by temples at virtually every fishing harbour. Other leading deities include Kwun Yum, the Buddhist goddess of mercy; Hung Shing, god of the South Seas and a weather prophet; and Wong Tai Sin, a Taoist saint and deity. The Christian group is estimated at about 500,000, comprising more than 50 Christian denominations and independent groups. The majority of the Christians are Roman Catholics, while the remainder belong to various Protestant sects such as Anglican, Baptist, Lutheran, and Methodist. There are also small numbers of Muslims, Hindus, Sikhs, and Jews.

Significant deities

Demographic trends. As in many large urban centres of the world, Hong Kong's population has increased in the late 20th century. Since the 1950s the average annual rate of growth has fluctuated between about 2 and 4 percent, the variations in some degree based on the sporadic flow of immigrants from China. Immigration has been a chief cause of population increase, but it was slowed through changes in immigration policy in 1980 and 1982. Birth and death rates have been steadily declining since the late 1950s, the rate of natural increase falling below 1 percent by the 1980s. Life expectancy, however, has been showing a gradual increase. Since the 1950s, the proportion of the population under 15 has decreased rapidly, while that between 15 and 64 has shown a marked increase and the group over 65 has more than doubled. Hong Kong is one of the world's most densely populated places.

THE ECONOMY

With its limited natural resources, Hong Kong depends on imports for virtually all of its requirements, including raw materials, food and other consumer goods, capital goods, and fuel. Under its unique status as an international free port, entrepôt trade, mainly with China, flourished until 1951, when a United Nations embargo on trade with China and North Korea drastically strangled it. This situation, combined with the need to export and with the availability of cheap labour, led to the establishment of competitive light industries and a transformation of the economy in the early 1960s. The market economy and the laissez-faire policy of the government provided flexibility for further industrialization and the incentive and freedom, since the late 1960s, to attract foreign investment and financial transactions. In succeeding years, with China adopting a more open foreign policy, entrepôt trade rapidly revived, while Hong Kong–China trade surged. Hong Kong has developed not only in manufacturing, trade, and shipping, but also as a regional financial centre and as an agent in China's pursuit of modernization. The tertiary (services) sector of the economy makes up more than one-half the gross domestic product (GDP), while manufacturing accounts for about one-fourth.

Stimulus of trade embargo

Resources. Hong Kong is practically devoid of any significant mineral resources. The mining for graphite and lead at Cham Shan ("Needle Hill") and iron ore at Ma On Shan stopped long ago. Small quantities of feldspar, feldspar sand, and kaolin clay are produced annually, mainly for domestic consumption. Hong Kong is similarly poorly endowed in other natural resources: no commercial timber is produced from its sparse forest cover, and there is no hydroelectric potential from the small and short streams. Indeed, even water has been in serious short supply as a consequence of the limited areal extent, the steep terrain, and the lack of catchment areas. In spite of the many reservoirs, which were built mostly before World War II, and several giant projects, such as the desalting plant at Castle Peak and the Plover Cove and the High Island reservoirs, which are enclosed sea areas, an increasing proportion of water consumed depends on supplies from China.

Lack of natural resources

Agriculture and fishing. Only 7 percent of Hong Kong's land area is arable, and another 2 percent is under fishponds. Since the 1950s almost 30 percent of the agricultural land has been lost to other uses. The growing season is year-round, however, and several crops a year are possible. Paddy rice cultivation once dominated agricultural land use, but it has practically disappeared, having been surpassed by vegetable and pond fish farming. Other minor uses include the production of fruits, flowers, and crops such as sweet potatoes, taro, yams, and sugarcane. About 40 percent of the farmland is abandoned or fallow.

Marine fishing in the adjacent waters is one of Hong Kong's most important primary activities. Apart from pond fish, a marine fish culture has shown signs of development, notably in the eastern New Territories.

Industry. The rapid industrial development in the 1950s was made possible by immigrant Chinese industrialists, mainly from Shanghai, who brought with them technology and capital. Foreign investments soon began to flow in to

Bob Davis—Aspect Picture Library, London



Vegetable farm on Pok Toi Chau (Lamma Island).

tap the huge supply of cheap labour and relatively cheap raw materials available in the surrounding region. Most of the industry has been confined to the urban areas, especially in the densely populated districts of Kowloon and New Kowloon. With the development of industrial and other new towns, industry began to disperse into Kwun Tong, Tsuen Wan, Tuen Mun, and others. In March 1977, the Hong Kong Industrial Estates Corporation was established to develop and manage industrial estates that would accommodate high-technology industries, first on reclaimed land in Tai Po and later in Yuen Long.

Manufacturing has become the most important sector of the Hong Kong economy. Although its percentage share of employment has dropped, it still employs almost 40 percent of the labour force. Textile and clothing production is the leading manufacturing activity, providing employment for more than 40 percent of the industrial labourers and contributing some 40 percent of the value of domestic exports. The electronics industry, the second largest export earner, and industries producing plastics, toys, and watches and clocks have displayed remarkable growth. Also developing are some heavy industries such as shipbuilding and repair and aircraft engineering, which enjoys an international reputation. Steel rolling, production of machine tools and parts, and cement manufacturing serve local needs. The tourist industry, which is highly promoted by the government and well catered to by the huge service sector, is another growing part of the economy. The number of tourists has topped 3,000,000 per year, and the industry generates almost 7 percent of the GNP.

Manufactured products

Finance. Since 1969, Hong Kong has been developing as a major financial centre of the Asia-Pacific region, despite the fact that it is without the services of a central bank. The government delegates the functions of such an institution to certain government offices and selected commercial banks. In addition to the licensed banks in the region, there are representative offices of foreign banks, including registered deposit-taking companies.

Domestic and international currencies are traded at the Hong Kong foreign-exchange market. The stock market attracts investment from both foreign and domestic sources. Some of its major shares are also traded on the London stock market. A gold bullion market, once the world's largest, is operated by the Chinese Gold and Silver Exchange Society. The lack of exchange controls has contributed to the success of Hong Kong as a financial centre.

Trade. Hong Kong's free-trade policy has made the territory one of the world's great centres of trade. There is no tariff on imports, except for some luxury items, such as perfumes, cars, alcoholic beverages, and tobacco. Hong Kong is dependent upon imported products, which make up about one-half the total amount of external trade, the rest being divided between exports and reexports. China and Japan supply the largest percentage of imports, and other major suppliers include the United States, Taiwan, Singapore, the United Kingdom, South Korea, and West Germany. Imports of raw materials and semimanufactured goods, such as electronic parts, man-made and natural textiles, and iron and steel represent the largest group of imports. The second largest group includes consumer goods such as clothing, radios, television sets, stereos, and diamonds and watches. Other significant imported products are machinery and transport equipment, foodstuffs, and fuels.

Imports

The United States provides the main market for the territory's products, other major markets being China, the United Kingdom, West Germany, Japan, and Canada. Textiles and clothing are the leading exports, followed by articles such as plastic toys, jewelry, and artificial flowers; electrical machinery and appliances; telecommunication and recording equipment; and a variety of other manufactured items and foods. Reexports constitute a major portion of the goods shipped out of Hong Kong, which developed historically as an entrepôt port.

Transportation. With roadways limited relative to the population, the government has enforced strict limitations on automobile ownership and placed heavy emphasis on the development of public transportation. As a result, car ownership is extremely low even when compared to other Asian cities. The majority of the populace makes its daily trips by public transport. Apart from the bus, tram (streetcar), and ferry, the public is also served by a unique minibus service, a rapid transit system, and an electric railway. Buses, however, are the largest carrier, responsible for more than half of the daily public transport trips excluding those by taxi, followed by the combined minibus and maxicab (a regulated form of minibus) service. The precipitous Victoria Peak area is served by one of the oldest transport companies, which operates a cable car system between the peak and the Central District.

Public transport

International traffic is well served by Hong Kong's international airport, its magnificent harbour, and overland linkages with China. The Hong Kong International Airport at Kai Tak is located on the eastern fringe of Kowloon. The port of Hong Kong, based at one of the world's finest natural harbours, is well known for its efficiency and capacity. The capacity of its container terminals at Kwai Chung ranks Hong Kong among the world's largest container ports. Speedy ferry service between Hong Kong and Macau and parts of China is provided by various craft, including hydrofoils and hovercraft. Railroad transportation to China is provided by the Kowloon–Canton Railway. Electrification of the railway and the growth along its line of the new towns of Sha Tin, Tai Po, and Fanling caused a considerable increase in passenger traffic. Externally, the line carries annually millions of tons of freight and head of livestock, as well as passenger traffic between Hong Kong and China.

GOVERNMENT AND SOCIAL CONDITIONS

Government. As a colony of the United Kingdom, Hong Kong's administration follows the pattern of other British-controlled territories. The governor, who is appointed by and represents the monarch of the United Kingdom, directs the government, serves as the commander in chief, and presides over the two main organs of government, the Executive Council and the Legislative Council. The Executive Council consists of four ex officio members and 12 appointed members, of whom two are official and 10 unofficial; they are appointed by the governor for fixed periods. The governor is required to seek the Executive Council's advice on all important matters. The Legislative Council comprises official members, including the governor and ex officio members, and unofficial members. None of the members of either council is elected, while most of the official and some of the unofficial memberships overlap. To develop representative government in Hong Kong, changes in the composition of the Legislative Council were scheduled to take effect in 1985.

An Urban Council, half elected and half appointed, is responsible for such services as street cleaning, refuse collection, health, and environmental control in the urban areas. In the New Territories, a Regional Council performs functions similar to those of the Urban Council. Each district board in the New Territories elects a representative to the council, and the Heung Yee Kuk (a consultative body of the New Territories) provides three ex officio members.

Local government

Since 1982, district boards and district management committees have been set up in each of Hong Kong's 18 administrative districts to encourage better communication with the public.

The law of Hong Kong is derived generally from that of the United Kingdom. Under certain special circumstances laws may be formulated for Hong Kong by order of the British monarch. The judiciary is headed by the chief justice, who is appointed by the governor on instructions from the British monarch, as are the justices of the Court of Appeal and the High Court. The governor also appoints district judges and the magistrates. The highest court in Hong Kong is the Court of Appeal, which, if required by the circumstances of the case, carries appeals to the Judicial Committee of the Privy Council in London for a final decision.

Education. Some 60 percent of the schools from kindergarten to secondary are private and about another 30 percent are either subsidized or aided by public funds. The number of public schools in Hong Kong is quite small. Education is compulsory through the junior sec-

ondary level. Students finishing primary, junior, and senior secondary education take examinations for allocation of school places at the next higher level.

Higher education

Postsecondary schools, mostly subsidized, are particularly strained by their small size. The combined enrollment of the two universities, the University of Hong Kong (1911) and the Chinese University of Hong Kong (1963), amounts to only about 14,000. The Hong Kong Polytechnic (1972), with mainly technical and vocational courses, and the Hong Kong Baptist College (1956), a thriving private institution, offer degree courses in selected subjects. Colleges of education train teachers, mainly for primary schools, while responsibility for teacher training for secondary education rests with the two universities. In 1984 the City Polytechnic of Hong Kong accepted its first students while its permanent campus was still being completed. Numerous other vocational, technical, and industrial training institutions operate throughout Hong Kong, and thousands of students are also enrolled in extension programs of the two universities and the Hong Kong Baptist College. About 80 percent of Hong Kong's people have at least a primary school education.

Housing. Historically, housing has been a major problem in Hong Kong, where space is limited and the number of occupants ever-growing. Changes in the residential environment between the establishment of the colony in 1842 and the Japanese occupation in 1941 were moderate, compared to those that took place in the postwar years. There was no planning in the earlier days of development, except that generally the British lived on the Peak (the area around Victoria Peak), other nationalities in the Mid-Levels (below the Peak), and the rich on somewhat higher ground, where the grand garden houses and big mansions remain as landmarks. Most of the Chinese lived on the lowlands surrounding the harbour, where the streets were narrow and the houses made of wood, bricks, and mortar. They lacked not only good natural lighting and ventilation but also piped water and flush toilets. Frequently urban development was the result of plagues, fires, and typhoons rather than comprehensive city planning; and the presence of large numbers of squatters and street sleepers as a result of a shortage of housing has been a persistent feature.

Public housing

The limited housing supply was further reduced by the ravages of World War II. In the early postwar years, more than half of all families shared accommodations with others, living in cubicles, bed spaces, and attics and on roofs and verandas and in similar quarters. The government's reluctant involvement in housing provision began with the building of resettlement blocks for fire victims in 1953, but it took real impetus in the early 1960s when the great demand for urban land resulted in the relocation of large numbers of squatters and urban poor. Eventually public housing came to accommodate some 45 percent of the population, most of them living far from the urban core, and with growing numbers settling into the new towns. Land prices and rents have reached extremely high levels.

Health and welfare. The health of the populace is generally good, the result, in part, of an aggressive program of public measures, including the promotion of preventive medicine and personal health services, and a relatively high quality of life. Improving health indexes and a downward trend in the occurrence of major communicable diseases are leading indicators of the state of health in Hong Kong. Most deaths are caused by cancer, heart disease, and cerebrovascular diseases. Hospitals are divided into three groups: government, government-assisted, and private. These are under great pressure to meet the needs of the people. Clinics, some operated by the government, supplement other medical facilities. Boat-borne clinics provide services to some outlying villages.

There is no social security system, except for a few emergency relief programs. The aging of the population, coupled with the extreme crowding, presents problems of elderly care that are increasing in scale and magnitude.

CULTURAL LIFE

Cultural milieu and the arts. Hong Kong's is truly a mixed culture. Not only does the territory celebrate festivals and holidays of the East and the West, such as the Dragon Boat Festival, the Mid-Autumn Festival, the Lunar (Chinese) New Year, Christmas, the Western New Year, and others, but it also enjoys hundreds of annual cultural events, ranging from traditional Cantonese and other Chinese regional operas and puppet shows to performances of ballet, theatre, and music and exhibitions of paintings and sculptures by nationally and internationally renowned performers and artists. The Hong Kong Arts Festival has become one of Asia's major cultural events, and the Hong Kong Philharmonic Orchestra, the Hong Kong Chinese Orchestra, the Chung Ying Theatre Company, and the City Contemporary Dance Company are among the best known local artistic groups. The Hong Kong Conservatory of Music and the Hong Kong Academy of Ballet have been combined into the Hong Kong Academy for Performing Arts, offering full-time diploma courses in dance, drama, music, and technical arts.

Music and theatre

Scores of motion pictures are produced every year in Hong Kong, many of which attain international fame; some have even started new trends in the art, such as the so-called kung fu films. The Hong Kong International Film Festival, inaugurated in 1977, is a major event, especially for the display of Asian films. Hong Kong is also a regional as well as an international centre in fashion design and in the cutting and design of ornamental diamonds.

Cultural institutions. Apart from the libraries of the major educational institutions, the Urban Council operates 25 libraries, including two mobile ones. Of the museums, major ones include those specializing in history, art, science and technology, and space. The multifunction City Hall (a cultural centre) and the Art Centre provide the major gallery, theatrical, and concert facilities. In addition, town halls have been established in the new towns and cultural centres in some districts to serve local communities.

Recreation. Hong Kong's country park system covers 40 percent of the land area, and outdoor recreation in parks is a part of the way of life for many of the people. City dwellers use park areas on the urban fringe for walking, running, and T'ai Chi ch'uan, among other activities, while remoter locations are used for kite flying, picnicking, hiking, cycling, and camping. The Urban Council and the Recreation and Sport Service are instrumental in providing organized programs of recreation and sports at the community level. The Ocean Park, called the world's largest oceanarium, the Hong Kong Coliseum, a 12,500-seat indoor stadium that is among the largest in Asia, and the Queen Elizabeth Stadium are among the best venues for local and international sports events and musical, cultural, and entertainment programs. For those who can afford it, the many inlets and bays in Hong Kong provide a superb setting for pleasure sailing, waterskiing, canoeing, and other water sports; the Hong Kong Yacht Club is one of the most active in the South China Sea region.

Press and broadcasting. A wide-ranging and sophisticated communications network has developed in Hong Kong, reflecting its thriving commerce and international importance. There are more than 60 newspapers (in various languages, but mostly Chinese) and the numbers of periodicals run into the hundreds. The territory is in addition the Far Eastern and Southeast Asian headquarters for most of the major international news services. Broadcast news is provided by several television and radio companies, one of which is government-run. There is no government control of the press other than that required to protect against libel and pornography. Television provides the major source of news and entertainment for the average family, and the Chinese television programs produced are not only for local consumption but also for overseas markets. Hong Kong also ranks as an important centre of publishing and printing: numerous books are published yearly for local consumption, several leading foreign publishers have their regional offices in Hong Kong, and many international magazines are printed in the territory.

History

EARLY SETTLEMENT

Archaeological remains of pottery, stone implements, rings, and bronzes found on more than 20 sites are evidence of

settlements in Neolithic times. The earliest modern peoples in Hong Kong are thought to have come from North China in the second millennium BC. The Cantonese began to settle in the area about 100 BC; later came the Hakka, and by the mid-17th century the Hoklo. Hong Kong was the scene of the last struggles between the declining Ming dynasty and the rising Ch'ing, led by the Manchus.

Before the British arrived in the mid-19th century, Hong Kong Island was inhabited only by a small fishing population, with few features to recommend it for settlement. It lacked fertile soil and fresh water, was mountainous, and was reputed to be a notorious haunt of pirates. But it was a relatively safe and undisturbed base for the British merchants who, in 1821, began to use the fine harbour to anchor opium-carrying vessels. The great commercial and strategic significance of this deep, sheltered harbour, possessing east and west entrances and lying on the main trade routes of the Far East, was quickly realized.

Arrival of the British

After the first Opium War (1839–42), Hong Kong Island was ceded to Britain by the Treaty of Nanking. The British were never satisfied with an incomplete control of the harbour, however. Less than 20 years later, after the second Opium War (1856–60), China was forced to cede Kowloon Peninsula south of what is now Boundary Street and Stonecutters Island by the Convention of Peking (1860). By the Convention of 1898, the New Territories together with 235 islands were leased to Britain for 99 years from July 1, 1898. With this expansion of territory, Hong Kong's population leaped to 120,000 in 1861 and to more than 300,000 by the end of the century.

EVENTS BEFORE AND DURING WORLD WAR II

Almost since its establishment, Hong Kong, more than any other treaty port, afforded a refuge for runaway persons and capital from China as well as an interim abode for rural emigrants destined for Southeast Asia and beyond. Such movements of Chinese people between China and Hong Kong were free and were highly responsive to the political and economic conditions in China. After the establishment of the Republic of China in 1912, proponents of emerging nationalism sought to abolish all foreign treaty privileges in China. A boycott against foreign goods particularly hurt Britain, which was well established in China. The campaign soon spread to Hong Kong, where strikes in the 1920s caused agitation.

When the Sino-Japanese War broke out in 1937, Hong Kong was once more a refuge, with thousands of Chinese fleeing to it before the advancing Japanese. With the outbreak of war in Europe in 1939, the position of the colony became even more precarious, and the Japanese attacked and occupied Hong Kong in December 1941. During the war years Hong Kong's commerce was drastically impaired; food was scarce, and many residents fled to inland China. The population, which numbered 1,600,000 in 1941, was reduced to about 650,000 by 1945 when the Japanese surrendered.

Japanese occupation

MODERN HONG KONG

British troops returned to the city on Aug. 30, 1945, and civil government was reestablished in May 1946. Meanwhile, hundreds of thousands of Chinese and foreigners returned, and they were soon joined by economic and political refugees from China, who were fleeing the civil war between the Nationalist and Communist armies.

The United Nations embargo in 1950 on trade with China and North Korea during the Korean War seriously depressed the entrepôt trade, the lifeline of the colony, and for several years conditions were depressed. Hong Kong began its revival based on light industries such as textiles, which were set up by immigrant capitalists and provided needed employment. These soon assumed their importance in the economy, providing as well the basis for further industrialization. But it was because much of the development depended on cheap labour, which toiled under extremely poor working conditions, that labour disputes and social discontent began to spread in the early 1960s. Severe riots broke out in Hong Kong and Kowloon in May 1967 following a labour dispute in a plastic-flower factory. The economic and social unrest was immediately turned into violent political demonstrations, largely inspired by followers of the Great Proletarian Cultural Revolution in China. When the situation stabilized toward the end of the 1960s, general working and living conditions were notably improved by labour legislation, large government housing projects, and extensive public works programs. Simultaneously, high-technology industries, such as electronics, were developed, and the property and financial markets prospered until early 1973, when the stock market collapsed as billions of dollars were drained out of Hong Kong. From the mid-1970s the economy resumed its upward trend as relations with China improved.

In the late 1970s, concern about the future of Hong Kong began to loom large, as British jurisdiction over the leased areas of the New Territories neared the 1997 expiration date. Although the lease applied only to the New Territories, the Chinese government had consistently maintained that the whole of Hong Kong was Chinese territory and considered that the question of the earlier Hong Kong–British agreements came under the category of unequal treaties and also required resolution. Initial contacts between the two governments on the matter were made from March 1979, but formal negotiations did not start until after the visit of the British prime minister to Peking in September 1982. Negotiations continued for two years. Finally, the Chinese-British joint declaration on the question of Hong Kong was formally signed by the heads of the two governments in Peking on Dec. 19, 1984. The agreement stipulated that Hong Kong (including Hong Kong Island, Kowloon, and the New Territories) would be recovered by China from July 1, 1997, and that upon resuming sovereignty over Hong Kong, the Chinese government, in accordance with the constitution, would establish a Hong Kong Special Administrative Region (HKSAR) directly under the central government. The agreement also stated that HKSAR would enjoy a high degree of autonomy, except in foreign affairs and defense, and that the current social and economic systems as well as the life-style in Hong Kong would remain unchanged for 50 years after 1997. It provided that these basic policies, to be stipulated in a basic law of HKSAR, would remain unchanged.

China–Britain agreement

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* especially sections 96/10 and 975, and the *Index.*

For later developments in the history of Hong Kong, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

BIBLIOGRAPHY. For geologic studies of Hong Kong, see P.M. ALLEN and E.A. STEPHENS, *Report on the Geological Survey of Hong Kong, 1967–1969* (1971); and BRYAN P. RUXTON, "The Geology of Hong Kong," *Quarterly Journal of the Geological Society of London,* 115:233–260 (1960). An analysis of the political, economic, geographic, and social developments in Hong Kong up to the early 1980s is found in CHI-KEUNG LEUNG, J.W. CUSHMAN, and WANG GUNGWU (eds.), *Hong Kong: Dilemmas of Growth* (1980); and a well-illustrated discussion of the geography, history, economy, and society is presented in *Hong Kong* (annual), issued by the Hong Kong Government Information Service. A compendium of statistical data is *Monthly Digest of Statistics,* issued by the Hong Kong Census and Statistics Department. FRANK LEEMING, *Street Studies in Hong Kong: Localities in a Chinese City* (1977), examines Hong Kong's neighbourhoods. Socioeconomic studies include *Hong Kong Social and Economic Trends, 1970–1980* (1981), compiled by the Hong Kong Census and Statistics Department; A.J. YOUNGSON, *Hong Kong, Economic Growth and Policy* (1982); SEK HONG NG and DAVID A. LEVIN, *Contemporary Issues in Hong Kong Labour Relations* (1983); E.F. SZCZEPANIK, *The Economic Growth of Hong Kong* (1958; reprinted with corrections, 1960); WILLIAM F. BEAZER, *The Commercial Future of Hong Kong* (1978); and THOMAS R. TREGEAR, *A Survey of Land Use in Hong Kong and the New Territories* (1958). *Hong Kong in Search of a Future* (1984), ed. by JOSEPH Y.S. CHENG; PETER HARRIS, *Hong Kong: A Study in Bureaucratic Politics* (1978); SIU-KAI LAU (CHAO-CHIA LIU), *Society and Politics in Hong Kong* (1982); and NORMAN MINERS, *The Government and Politics of Hong Kong* (3rd ed., 1981), address the politics of the colony. See also DAVID FU-KEUNG IP, CHI-KEUNG LEUNG, and CHUNG-TONG WU (comps.), *Hong Kong: A Social Sciences Bibliography* (1974); and ALAN BIRCH, Y.C. JAO, and ELIZABETH SINN (eds.), *Research Materials for Hong Kong Studies* (1984).

(C.-K.L.)

Horses and Horsemanship

Zoologically, the horse is a mammal of the family Equidae. It comprises a single species, *Equus caballus,* whose numerous varieties are called breeds. Before the advent of mechanized vehicles, riding on horseback was one of the chief means of transportation. Horsemanship evolved, of necessity, as the art of riding with maximum discernment and a minimum of interference with the horse. Until the 20th century riding was a monopoly of the cavalry, of cowboys and others whose work required riding on horseback, and of the wealthy, who rode for sport. Although hunting and polo tend to remain sports of the wealthy and the role of the horse in battle has ended, special value is now placed on horse shows of a high standard, in which the most popular event is undoubtedly show jumping. Horsemanship has remained a valued social asset and symbol of prestige, but the opening of many new riding clubs and stables has made riding and horsemanship accessible to a much larger segment of the population. (C.E.C./Ed.)

This article is divided into the following sections:

General features of horses 646
- Form and function 646
  - Anatomical adaptations
  - Senses
  - Colour and pattern
  - Nutrition
  - Behaviour
  - Reproduction and development
  - Diseases and parasites
- Breeds of horses 648
  - Light horses
  - Heavy breeds
  - Ponies
- Evolution of the horse 650

Riding and horsemanship 651
- History 651
  - Origins and early history
  - Military horsemanship
- The art of horsemanship 651
  - Fundamentals
  - The horse's movements
  - Training and equipment
  - Seats
  - Dressage
  - Jumping
- Riding and shows 654
  - Horse shows
  - Olympic equestrian competition

Bibliography 655

General features of horses

Utility of the wild horse

In prehistoric times the wild horse was probably first hunted for food. When its domestication took place is unknown, but it certainly was long after the domestication of the dog or of cattle. It is supposed that the horse was first used by a tribe of Indo-European origin that lived in the steppes north of the chain of mountains adjacent to the Black and Caspian seas. Influenced by climate, food, and humans, the horse rapidly acquired its present form.

The relationship of the horse to humans has been unique. The horse is a partner and friend. It plowed fields and brought in the harvest, hauled goods and conveyed passengers, followed game and tracked cattle, and carried combatants into battle and adventurers to unknown lands. It has provided recreation in the form of jousts, tournaments, carousels, and the sport of riding. The influence of the horse is expressed in the English language in such terms as *chivalry* and *cavalier,* which connote honour, respect, good manners, and straightforwardness.

The horse is the "proudest conquest of Man," according to the French zoologist Le Comte de Buffon. Its place was at its master's side in the graves of the Scythian kings or in the tombs of the pharaohs. Many early human cultures were centred on possession of the horse. Superstition read meaning into the colours of the horse, and a horse's head suspended near a grave or sanctuary or on the gables of a house conferred supernatural powers on the place. Greek mythology created the centaur, the most obvious symbol of the oneness of horse and rider. White stallions were the supreme sacrifice to the gods, and the Greek general Xenophon recorded that "Gods and heroes are depicted on well-trained horses." A beautiful and well-trained horse was, therefore, a status symbol in ancient Greece. Kings, generals, and statesmen, of necessity, had to be horsemen. The names of famous horses are inseparably linked to those of their famous riders: Bucephalus, the charger of Alexander the Great; Incitatus, foolishly made a senator by the Roman emperor Caligula; El Morzillo, Cortés' favourite horse, to whom the Indians erected a statue; Roan Barbery, the stallion of Richard II, mentioned by Shakespeare; Copenhagen, the Duke of Wellington's horse, which was buried with military honours.

The horse in art

The horse has occupied a special place in the realm of art. From the Stone Age drawings to the marvel of the Parthenon frieze, from Chinese T'ang dynasty tomb sculptures to Leonardo da Vinci's sketches and Verrocchio's Colleoni, from the Qur'ān to modern literature, the horse has inspired artists of all ages and in all parts of the world.

The horse in life has served its master in travels, wars, and labours and in death has provided many commodities. Long before their domestication horses were hunted by primitive tribes for their flesh, and horsemeat is still consumed by people in parts of Europe and in Iceland and is the basis of many pet foods. Horse bones and cartilage are used to make glue. Tetanus antitoxin is obtained from the blood serum of horses previously inoculated with tetanus toxoid. From horsehide a number of articles are manufactured, including fine shoes and belts. The cordovan leather fabricated by the Moors in Córdoba, Spain, was originally made from horsehide. Stylish fur coats are made of the sleek coats of foals. Horsehair has wide use in upholstery, mattresses, and stiff lining for coats and suits; high-quality horsehair, usually white, is employed for violin bows. Horse manure, which today provides the basis for cultivation of mushrooms, was used by the Scythians for fuel. Mare's milk was drunk by the Scythians, the Mongols, and the Arabs.

FORM AND FUNCTION

A mature male horse is called a stallion, the female a mare. A stallion used for breeding is known as a stud. A castrated stallion is commonly called a gelding. Formerly, stallions were employed as riding horses, while mares were kept for breeding purposes only. Geldings were used for work and as ladies' riding horses. Recently, however, geldings generally have replaced stallions as riding horses. Young horses are known as foals; male foals are called colts and females fillies.

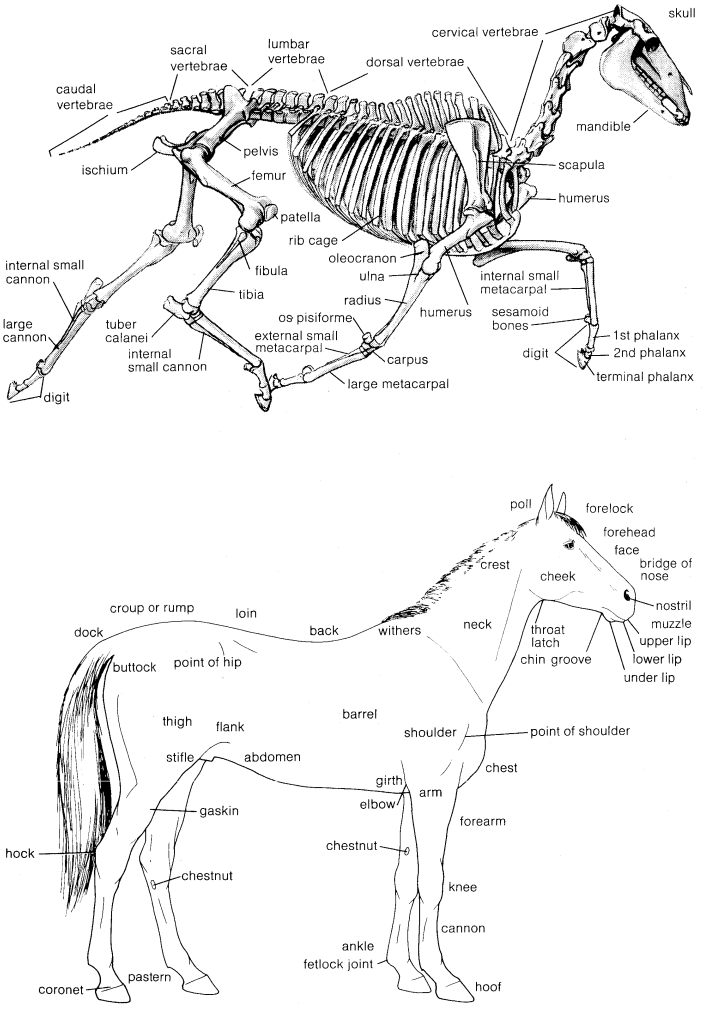
Anatomical adaptations. The primitive horse probably stood 12 hands tall (about 120 centimetres, or 48 inches) at the withers, the high point on the back at the base of the neck, and was dun coloured. Domestic horses gone wild, such as the mustangs of western North America, tend to revert to those primitive features under random mating; they generally are somewhat taller (about 15 hands), usu-

ally gray, dun, or brownish in colour, and move in herds led by a stallion.

The horse's general form is characteristic of an animal of speed: the long leg bones pivot on pulley-like joints that restrict movement to the fore and aft, the limbs are levered to muscle masses in such a way as to provide the most efficient use of energy, and the compact body is supported permanently on the tips of the toes, allowing fuller extension of the limbs in running (see illustration).

The rounded skull houses a large and complex brain, well developed in those areas that direct muscle coordination. While the horse is intelligent among subhuman animals, it is safe to say that the horse is more concerned with the functioning of its acute sensory reception and its musculature than with mental processes. Though much has been written about "educated" horses that appear to exhibit an ability to spell and count, it is generally agreed that in such cases a very perceptive animal is responding to cues from its master. But this ability is remarkable enough in its own right—for the cues are often given unconsciously by the human trainer, and detection of such subtle signals requires extremely sharp perception.

The horse, like other grazing herbivores, has typical adaptations for plant eating: a set of strong, high-crowned teeth, suited to grinding grasses and other harsh vegetation, and a relatively long digestive tract, most of which is intestine concerned with digesting cellulose matter from vegetation. Young horses have milk (or baby) teeth, which they begin to shed at about age two and a half. The permanent teeth, numbering 36 to 40, are completely developed by age four to five years. In the stallion these teeth are arranged as follows on the upper and lower jaws: 12 incisors that cut and pull at grasses; four canines, remnants without function in the modern horse and usually not found in mares; 12 premolars and 12 molars, high prisms that keep moving out of the jaw to replace the surfaces worn off in grinding food.



(Top) Skeleton of the horse; (bottom) points of the horse.

Under domestication the horse has diversified into three major types, based on size and build: draft horses, heavy limbed and up to 20 hands high (200 centimetres, or 80 inches); ponies, by convention horses under 14.2 hands high (about 144 centimetres, or 57 inches); and light horses—the saddle or riding horses—which fall in the intermediate size range. Domestic horses tend to be near-sighted, less hardy than their ancestors, and often high-strung, especially thoroughbreds, where intensive breeding has been focused upon speed to the exclusion of other qualities. The stomach is relatively small, and, since much vegetation must be ingested to maintain vital processes, foraging is almost constant under natural conditions. Domestic animals are fed several (at least three) times a day in quantities governed by the exertion of the horse.

Diversification into three types

Senses. The extremely large eyes placed far back on the elongated head admirably suit the horse for its chief mode of defense: flight. Its long neck and high-set eyes, which register a much wider range than do the eyes of a human being, enable the horse to discern a possible threat even while eating low grasses. Like human vision, the horse's vision is binocular, but only in the narrow area directly forward, and evidence suggests that it does not register colour. While visual acuity is high, the eyes do not have variable focus, and objects at different distances register only on different areas of the retina, which requires tilting movements of the head. The senses of smell and hearing seem to be keener than in human beings. As the biologist George Gaylord Simpson has put it in *Horses* (1961): "Legs for running and eyes for warning have enabled horses to survive through the ages, although subject to constant attack by flesh eaters that liked nothing better than horse for supper."

Colour and pattern. From the dun of the primitive horse has sprung a variety of colours and patterns, some highly variable and difficult to distinguish. Among the most important colours are black, bay, chestnut (and sorrel), palomino, cream, and white.

The black colour is a true black, although a white face marking (blaze) and white ankles (stockings) may occur. The brown horse is almost black but has lighter areas around the muzzle, eyes, and legs. Bay refers to several shades of brown, from red brown and tan to sandy. Bay horses have a black mane, tail, and (usually) stockings. Chestnut is similar to bay but with none of the bay's black overtones. Lighter shades of chestnut are called sorrel. The palomino horse runs from cream to bronze, with a flaxen or silvery mane and tail. The cream is a diluted sorrel, or very pale yellow, nearly white. White in horses is variable, ranging from aging grays (see below) to albinos with blue eyes and pink skin and to pseudoalbinos with a buff mane or with brown eyes. The chief patterns of the white horse are gray, roan, pinto, and appaloosa. Gray horses are born dark brown or black and develop white hairs as they age, becoming almost all white in advanced years. Roan refers to white mixed with other colours at birth: blue roan is white mixed with black; red roan is mixed white and bay; and strawberry roan is white and chestnut. The pinto is almost any spotted pattern of white and another colour; other names, such as paint, calico, piebald, skewbald, overo, and tobiano, refer to subtle distinctions in type of colour or pattern. Appaloosa is another extremely variable pattern, but the term generally refers to a large white patch over the hips and loin, with scattered irregular dark spots.

Nutrition. The horse's natural food is grass. For stabled horses, the diet generally consists of hay and grain. The animal should not be fed immediately before or after work, to avoid digestive problems. Fresh water is important, especially when the horse is shedding its winter coat, but the animal should never be watered when it is overheated after working. Oats provide the greatest nutritional value and are given especially to foals. Older horses, whose teeth are worn down, or those with digestive troubles, can be provided with crushed oats. Chaff (minced straw) can be added to the oat ration of animals that eat greedily or do not chew the grain properly. Crushed barley is sometimes

substituted in part for oats. Hay provides the bulk of the horse's ration and may be of varying composition according to locale. Mash is bran mixed with water and with various invigorating additions or medications. It may be given to horses with digestive troubles or deficient eating habits. Corn (maize) is used as a fattening cereal, but it makes the horse sweat easily. Salt is needed by the horse at all times and especially when shedding. Bread, carrots, and sugar are tidbits often used to reward an animal by the rider or trainer. In times of poverty horses have adapted to all sorts of food—potatoes, beans, green leaves, and in Iceland even fish—but such foods are not generally taken if other fare is available. A number of commercial feed mixes are available to modern breeders and owners; these mixes contain minerals, vitamins, and other nutrients and are designed to provide a balanced diet when supplemented with hay.

Behaviour. The horse's nervous system is highly developed and gives proof to varying degrees of the essential faculties that are the basis of intelligence: instinct, memory, and judgment. Foals, which stand on their feet a short while after birth and are able to follow their mothers within a few hours, even at this early stage in life exhibit the traits generally ascribed to horses. They have a tendency to flee danger. They express fear sometimes by showing panic and sometimes by immobility. Horses rarely attack and do so either when flight is impossible or when driven to assault a person who has treated them brutally.

Habit and instinct

Habit governs a large number of their reactions. Instinct, together with a fine sense of smell and hearing, enables them to sense water, fire, even distant danger. An extremely well-developed sense of direction permits the horse to find its way back to its stables even at night or after a prolonged absence. The visual memory of the horse prompts it to shy repeatedly from an object or place where it had earlier experienced fear. The animal's auditory memory, which enabled ancient army horses or hunters to follow the sounds of the bugles, is used in training. When teaching, the instructor always uses the same words and the same tone of voice for a given desired reaction. Intelligent horses soon attach certain movements desired by their trainers to particular sounds and even try to anticipate their rider's wishes.

While instinct is an unconscious reaction more or less present in all individuals of the same species, the degree of its expression varies according to the individual and its development. Most horses can sense a rider's uncertainty, nervousness, or fear and are thereby encouraged to disregard or even deliberately disobey the rider. Highbred animals, which give evidence of greater intelligence than those of low breeding, are capable not only of acts of vengeance and jealousy against their riders but also of expressions of confidence, obedience, affection, and fidelity. They are less willing than a lowbred horse to suffer rough handling or unjust treatment.

Cunning animals have been known to employ their intelligence and physical skill to a determined end, such as opening the latch of a stall or the lid of a chest of oats.

Reproduction and development. The onset of adult sex characteristics generally begins at the age of 16 to 18 months; the horse is considered mature, according to the breed, at approximately three years and adult at five. Fecundity varies according to the breed and may last beyond age 20 with thoroughbreds and to 12 or 15 with other horses. The gestation period is 11 months, 280 days being the minimum in which the foal can be born with expectation to live. As a rule a mare produces one foal per mating, twins occasionally, and triplets rarely. The foal is weaned at six months.

The useful life of a horse varies according to the amount of work it is required to do and the maintenance furnished by its owner. A horse that is trained carefully and slowly and is given the necessary time for development may be expected to serve to an older age than a horse that is rushed in its training. Racehorses that enter into races at the age of two rarely remain on the turf beyond eight. Well-kept riding horses, on the contrary, may be used more than 20 years.

Life expectancy

The life span of a horse is calculated at six to seven times the time necessary for his physical and mental development; that is, 30 to 35 years at the utmost, the rule being about 20 to 25 years. Ponies generally live longer than larger horses. There are a number of examples of horses that have passed the usual limit of age. The veterinary university of Vienna conserves the skeleton of a thoroughbred mare of 44 years of age. There have been reports made of horses living to the early 60s in age.

Diseases and parasites. Horses are subjected to a number of contagious diseases, such as influenza, strangles, glanders, equine encephalomyelitis, and swamp fever. Their skin is affected by parasites, including certain mites, ticks, and lice. Those with sensitive skin are especially subject to eczemas and abscesses, which may result from neglect or contamination. Sores caused by injuries to the skin from ill-fitting or unclean saddles and bridles are common ailments. The horse's digestive tract is particularly sensitive to spoiled feed, which causes acute or chronic indigestion, especially in hot weather. Worms can develop in the intestine and include the larvae of the botfly, pinworms, tapeworms, and roundworms (ascarids). Overwork and neglect may predispose the horse to pneumonia and rheumatism. The ailment known as roaring is an infection of the larynx that makes the horse inhale noisily; a milder form causes the horse to whistle. Chronic asthma, or "broken wind," is an ailment that is considered to be all but incurable. A horse's legs and feet are sensitive to blows, sprains, and overwork, especially if the horse is young or is worked on hard surfaces. Lameness may be caused by bony growths, such as splints, spavins, and ringbones, by soft-tissue enlargements, known as windgalls, thoroughpins, and shoe boils, and by injury to the hooves, including sand crack, split hoof, tread thrush, and acute or chronic laminitis.

BREEDS OF HORSES

The first intensively domesticated horses were developed in Central Asia. They were small, lightweight, and stocky. In time, two general groups of horses emerged: the southerly Arab-Barb types (from the Barbary coast) and the northerly, so-called cold-blooded types. When, where, and how these horses appeared is disputed. Nevertheless, all modern breeds—the light, fast, spirited breeds typified by the modern Arabian, the heavier, slower, and calmer working breeds typified by the Belgian, and the intermediate breeds typified by the Thoroughbred—may be classified according to where they originated (*e.g.,* Percheron, Clydesdale, and Arabian), by the principal use of the horse (riding, draft, coach horse), and by their outward appearance and size (light, heavy, pony).

Light horses. *Arabian.* Its long history is obscured by legend, but the Arabian breed, prized for its stamina, intelligence, and character, is known to have been developed in Arabia by the 7th century AD. It is a compact horse with a small head, protruding eyes, wide nostrils, marked withers, and a short back. It usually has only 23 vertebrae, while 24 is the usual number for other breeds. Its legs are strong with fine hooves. The coat, tail, and mane are of fine silky hair. While many colours are possible in the breed, gray prevails. The most famous stud farm is in the region of Najd, Saudi Arabia, but many fine Arabian horses are bred in the United States.

Thoroughbred lineage

Thoroughbred. The history of the English Thoroughbred is a long one. Records indicate that a stock of Arab and Barb horses was introduced into England as early as the 3rd century. Conditions of climate, soil, and water favoured development, and selective breeding was long encouraged by those interested in racing. Under the reigns of James I and Charles I, 43 mares (the Royal Mares) were imported into England, and a record, the General Stud Book, was begun in which are inscribed only those horses that may be traced back to the Royal Mares in direct line, or to only three other horses imported to England: the Byerly Turk (imported in 1689), the Darley Arabian (after 1700), and the Godolphin Barb (also known as the Godolphin Arabian, imported about 1730). The English Thoroughbred has since been introduced to most countries, where it is bred for racing or used to improve local breeds. The Thoroughbred has a small fine head, a

deep chest, and a straight back. Its legs have short bones that allow a long easy stride, and its coat is generally bay or chestnut, rarely black or gray.

Asian. Asian breeds were strongly influenced by Arabian or Persian breeds, which together with the horses of the steppes produced small, plain-looking horses of great intelligence and endurance. Among them are the Tartar, Kirghis, Mongol, and Cossack horses. A Persian stallion and a Dutch mare produced the Orlov trotter in 1778, named after Count Orlov, the owner of the stud farm where the mating took place.

Anglo-Arab. The Anglo-Arab breed originated in France with a crossing of English Thoroughbreds to pure Arabians. The matings produced a horse larger than the Arabian and smaller than the Thoroughbred, of easy maintenance, and capable of carrying considerable weight in the saddle. Its coat is generally chestnut or bay.

American breeds. The Standardbred, a breed that excels at the pace and trot, ranks as one of the world's finest harness racers. A powerful, long-bodied horse, the Standardbred was developed during the first half of the 19th century and can be traced largely to the sire Messenger, a Thoroughbred imported from Britain in 1788 and mated to various brood mares in New York, New Jersey, and Pennsylvania.

The quarter horse was bred for races of a quarter of a mile and is said to descend from Janus, a small Thoroughbred stallion imported into Virginia toward the end of the 18th century. It is 14.2 to 16 hands high, with sturdily muscled hindquarters, essential for the fast departure required in short races. It serves as a polo pony equally well as for ranch work.

The Morgan horse originated from a stallion given to Justin Morgan of Vermont around 1795. This breed has become a most versatile horse for riding, pulling carriages, farm labour, and cattle cutting. It was the ideal army charger. It stands about 15 hands high and is robust, good-natured, willing, and intelligent. Its coat is dark brown or liver chestnut.

The Appaloosa

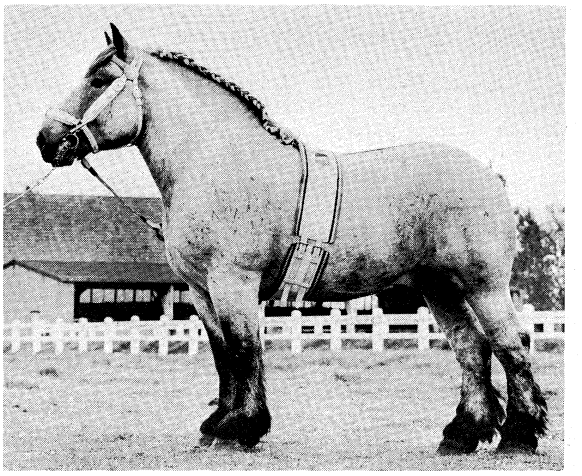
Appaloosa is a colour breed (see above) said to have descended in the Nez Percé Indian territory of North America from wild mustangs, which in turn descended from Spanish horses brought to the New World by explorers. The Appaloosa is 14.2 to 15.2 hands high, of sturdy build and of most diverse use; it is especially good in farm work.

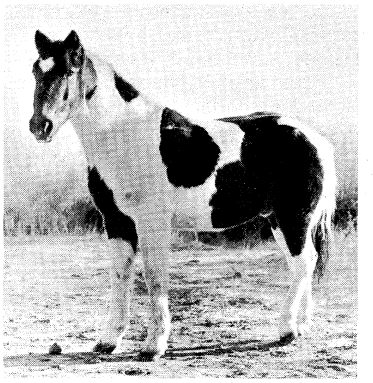
American breeders have also developed several horses that have specialized gaits. These gaited breeds include the American saddle horse, the Tennessee walking horse, and the Missouri fox trotting horse. The American saddle horse has a small head and spectacular high-stepping movements. It is trained for either three or five gaits. The three-gaited horses perform the walk, trot, and canter; the five-gaited horses in addition perform the rack, a quick, high-stepping four-beat gait, and the slow gait, a somewhat slower form of the rack. Since they are used mainly for shows, their hooves are kept rather long, and the muscles of the tail are often clipped so that the base of the tail is carried high. Chestnut and bay are the usual colours. The Tennessee walking horse—a breed derived partially from the Thoroughbred, Standardbred, Morgan, and American saddle horse—serves as a comfortable riding mount used to cover great distances at considerable speed. Its specialty is the running walk, a long and swift stride. Bay is the most common colour. The Missouri fox trotting horse, a breed developed to cover the rough terrain of the Ozark region, is characterized by an unusual gait, called the fox trot, in which the front legs move at a walk while the hind legs perform a trot. The most common colours for this breed are sorrel and chestnut sorrel.

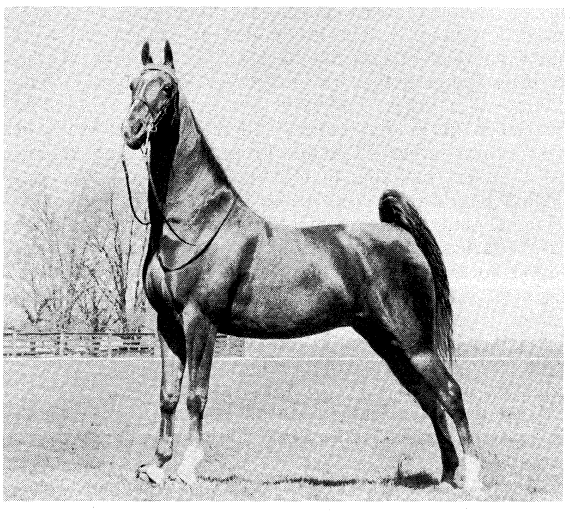
Other light breeds. The English Hackney is a light carriage horse, influenced by the Thoroughbred, capable of covering distances of 12 to 15 miles (19 to 24 kilometres) per hour at the trot and canter. It measures 15.2 to 15.3 hands high and is appreciated for its high knee action.

The Cleveland Bay carriage horse, up to 17 hands high and generally bay in colour, is similar to the Yorkshire Coach horse. Both breeds are now used for the sport of driving.

Other versatile breeds include the German Holstein,







Representative horse breeds.
(Top) The Belgian, a heavy breed; (centre) an Indian pony; (bottom) the American saddle horse, a light breed.
Sally Anne Thompson

Hanoverian, and East Prussian (Trakehner), which serve equally well for riding, light labour, and carriage. These horses, 16 to 18 hands high and of all colours, are now mostly bred for sport.

The Andalusian, a high-stepping, spirited horse, and the small but enduring Barb produced the Lipizzaner, which was named after the stud farm founded near Trieste, Italy, in 1580. Originally of all colours, the Lipizzaner is gray or, now exceptionally, bay. It is small, rarely over 15 hands high, of powerful build but slender legs, and with long silky mane and tail. Intelligence and sweetness of disposition as well as gracefulness destined it for academic horsemanship, notably as practiced at the Spanish Riding School of Vienna.

Heavy breeds. The horses used for heavy loads and

farm labour descended from the ancient war horses of the Middle Ages. These breeds—including the English Shire (the world's largest horse), Suffolk, and Clydesdale; the French Percheron; the Belgian horse; the German Noriker; and the Austrian Pinzgauer—are now little used for their original purpose, having been almost entirely replaced by the tractor. They usually measure well over 16 hands high, some more than 19 hands. They are of all colours, sometimes spotted, and generally have a very calm temperament.

Originating in the South Tyrol, the Haflinger is a mountain horse, enduring, robust, and versatile, used for all farm labour, for pulling a carriage or sledge, and for pack hauling. It is rarely over 14.2 hands high and is chestnut with a flaxen mane and tail.

Ponies. Ponies are any horses other than Arabians that are shorter than 14.2 hands. They are generally very sturdy, intelligent, energetic, and sometimes stubborn. The coat is of all colours, mainly dark, and the mane and tail are full. Ponies are used for pulling carriages and pack loads and as children's riding horses or pets. There are numerous varieties, including the Welsh, Dartmoor, Exmoor, Connemara, New Forest, Highland, Dale, Fell, pony of the Americas, Shetland (under seven hands high), Iceland, and Norwegian. Ponies of the warmer countries include the Indian, Java, Manila, and Argentine. (A.W.P./Ed.)

EVOLUTION OF THE HORSE

The evolutionary lineage of the horse is among the best documented in all paleontology. The history of the horse family, Equidae, began during the Eocene Epoch, which lasted from about 54,000,000 to 38,000,000 years ago. During the early Eocene there appeared the first ancestral horse, a hoofed, browsing mammal known technically as *Hyracotherium* but more commonly called eohippus, the "dawn horse." Fossils of eohippus, which have been found in both North America and Europe, show an animal that stood from 4.2 to 5 hands high, diminutive by comparison to the modern horse, with an arched back and raised hindquarters. The legs ended in padded feet with four functional hooves on the forefeet and three on the hindfeet—quite unlike the unpadded, single-hoofed foot of modern equines. The skull lacked the large, flexible muzzle of the modern horse, and the size and shape of the cranium indicate that the brain was far smaller and less complex than that of today's horse. The teeth, too, differed significantly from those of the modern equines, being adapted to a fairly general browser's diet. Eohippus was, in fact, so unhorselike that its evolutionary relationship to the modern equines was at first unsuspected. It was not until paleontologists had unearthed fossils of later extinct horses that the link to eohippus became clear.

Trends in the evolution of the horse

The line leading from eohippus to the modern horse exhibits the following evolutionary trends: increase in size, reduction in the number of hooves, loss of the foot pads, lengthening of the legs, fusion of the independent bones of the lower legs, elongation of the muzzle, increase in the size and complexity of the brain, and the development of crested, high-crowned teeth suited to grazing. This is not to imply that there was a steady, gradual progression in these characteristics leading inevitably from those of eohippus to those of the modern horse. Some of these features, such as grazing dentition, appear abruptly in the fossil record, rather than as the culmination of numerous, gradual changes. Eohippus, moreover, gave rise to many now-extinct branches of the horse family, some of which differed substantially from the line leading to the modern equines.

Although eohippus fossils occur in both the Old and New worlds, the subsequent evolution of the horse took place chiefly in North America. During the remainder of the Eocene, the prime evolutionary changes were in dentition. *Orohippus,* a genus from the middle Eocene, and *Epihippus,* a genus from the late Eocene, resembled eohippus in size and in the structure of the limbs. But the form of the cheek teeth—the four premolars and the three molars found in each half of both jaws—had changed somewhat. In eohippus the premolars and molars were clearly distinct, the molars being larger. In *Orohippus* the fourth premolar had become similar to the molars, and in *Epihippus* both the third and fourth premolars had become molar-like. In addition, the individual cusps that characterized the cheek teeth of eohippus had given way in *Epihippus* to a system of continuous crests or ridges running the length of the molars and molariform premolars. These changes, which represented adaptations to a more specialized browsing diet, were retained by all subsequent ancestors of the modern horse.

Fossils of *Mesohippus,* the next important ancestor of the modern horse, are found in the early and middle Oligocene of North America (the Oligocene Epoch lasted from about 38,000,000 to 26,000,000 years ago). *Mesohippus* was far more horselike than its Eocene ancestors—it was larger (averaging about six hands high); the snout was more muzzlelike; and the legs were longer and more slender. *Mesohippus* also had a larger brain. The fourth toe on the forefoot had been reduced to a vestige, so that both the forefeet and hindfeet carried three functional toes and a foot pad. The teeth remained adapted to browsing.

By the late Oligocene, *Mesohippus* had evolved into a somewhat larger form known as *Miohippus.* The descendants of *Miohippus* split into various evolutionary branches during the early Miocene (the Miocene Epoch lasted from about 26,000,000 to 7,000,000 years ago). One of these branches, known as the anchitheres, included a variety of three-toed, browsing horses comprising several genera. Anchitheres were successful, and some genera spread from North America across the Bering land bridge into Eurasia.

It was a different branch, however, that led from *Miohippus* to the modern horse. The first representative of this line, *Parahippus,* appeared in the early Miocene. *Parahippus* and its descendants marked a radical departure in that they had teeth adapted to eating grass. Grasses were at this time becoming widespread across the North American plains, providing *Parahippus* with a vast food supply.

Dental adaptations to a grass diet

Grass is a much coarser food than succulent leaves and requires a different kind of tooth structure. The cheek teeth developed larger, stronger crests and became adapted to the side-to-side motion of the lower jaw necessary to grind grass blades. Each tooth also had an extremely long crown, most of which, in the young animal, was buried beneath the gum line. As grinding wore down the exposed surface, some of the buried crown grew out. This high-crowned tooth structure assured the animal of having an adequate grinding surface throughout its normal life span. Adaptations in the digestive tract must have occurred as well, but the organs of digestion are not preserved in the fossil record.

The change from browsing to grazing dentition was essentially completed in *Merychippus,* which evolved from *Parahippus* during the middle and late Miocene. *Merychippus* must have looked much like a modern pony. It was fairly large, standing about 10 hands high, and its skull was similar to that of the modern horse. The long bones of the lower leg had become fused; this structure, which has been preserved in all modern equines, is an adaptation for swift running. The feet remained three-toed, but in many species the foot pad was lost, and the two side toes became rather small. In these forms, the large central toe bore the animal's weight. Strong ligaments attached this hoofed central toe to the bones of the ankles and lower leg, providing a spring mechanism that pushed the flexed hoof forward after the impact of hitting the ground. *Merychippus* gave rise to numerous evolutionary lines during the late Miocene. Most of these, including *Hipparion, Neohipparion,* and *Nannippus,* retained the three-toed foot of their ancestors. One line, however, led to the one-toed *Pliohippus,* the direct predecessor of *Equus. Pliohippus* fossils occur in the early to middle Pliocene beds of North America (the Pliocene Epoch lasted from about 7,000,000 to 2,500,000 years ago).

Equus—the genus to which all modern equines, including horses, asses, and zebras, belong—evolved from *Pliohippus* toward the end of the Pliocene. *Equus* shows even greater development of the spring mechanism in the foot and exhibits straighter and longer cheek teeth. This new form was extremely successful and had spread from

the plains of North America to South America and to all parts of the Old World by the early Pleistocene (the Pleistocene Epoch lasted from about 2,500,000 to 10,000 years ago). *Equus* flourished in its North American homeland throughout the Pleistocene; then, about 10,000 to 8,000 years ago, *Equus* disappeared from North and South America. Scholars have offered various explanations for this disappearance, including the emergence of devastating diseases or the arrival of human populations (which presumably hunted the horse for food). Despite these speculations, the reasons for the demise of *Equus* in the New World remain uncertain. The submergence of the Bering land bridge prevented any return migration of horses from Asia, and *Equus* was not reintroduced into its native continent until the Spanish explorers brought horses in the early 16th century.

The demise of *Equus* in the Americas

During the Pleistocene, the evolution of *Equus* in the Old World gave rise to all the modern members of the genus. The modern horse, *Equus caballus,* became widespread from central Asia to most of Europe. Local types of horses, all breeds of this single species, undoubtedly developed, and three of these—Przewalski's horse from central Asia, the tarpan from eastern Europe and the Ukrainian steppes, and the forest horse of northern Europe—are generally credited as being the ancestral stock of the domestic horse. According to this line of thinking, Przewalski's horse and the tarpan formed the basic breeding stock from which the southerly "warm-blooded" horses developed, while the forest horse gave rise to the heavy, "cold-blooded" breeds.

(Ed.)

Riding and horsemanship

HISTORY

Origins and early history. From the 2nd millennium BC, and probably even earlier, the horse was employed as a riding animal by fierce nomadic peoples of central Asia. One of these peoples, the Scythians, were accomplished horsemen and used saddles. It is also likely that they realized the importance of a firm seat and were the first to devise a form of stirrup. A saddled horse with straps hanging at the side and looped at the lower end is portrayed on a vase of the 4th century BC found at Chertomlyk in the Soviet Union. This contrivance may have been used for mounting only, however, because of the danger of being unable to free the foot quickly in dismounting. The Greek historian Strabo said that the indocility of the Scythians' wild horses made gelding necessary, a practice until then unknown in the ancient world. The Sarmatians, superb horsemen who superseded the Scythians, rode bareback, controlling their horses with knee pressure and distribution of the rider's weight.

Among the earliest peoples to fight and hunt on horseback were the Hittites, the Assyrians, and the Babylonians; at the same time (about 1500 BC) the Hyksos, or Shepherd Kings, introduced horses into Egypt and rode them in all their wars. In the 8th and 7th centuries BC, the Scythians brought horses to Greece, where the art of riding developed rapidly, at first only for pleasure. A frieze from the Parthenon in Athens shows Greeks riding bareback. Philip II of Macedon had a body of cavalry in his army, and the army of his son Alexander had separate, organized horse units. In the 4th century BC another Greek historian, Xenophon, wrote his treatise *Peri hippikēs* (*On Horsemanship*), giving excellent advice on horsemanship. Many of his principles are still perfectly valid. He advocated the use of the mildest possible bits and disapproved of the use of force in training and in riding. The Roman mounted troops were normally barbarian archers who rode without stirrups and apparently without reins, leaving the hands free to use the bow and arrow.

Classical horsemanship

As a general rule almost every item of riding equipment used today originated among the horsemen of the Eurasian steppes and was adopted by the people of the lands they overran to the east, the south, and later the west.

Horseshoes of various types were used by migratory Eurasian tribes about the 2nd century BC, but the nailed iron horseshoe as used today first appeared in Europe about the 5th century AD, introduced by invaders from the East. One, complete with nails, was found in the tomb of the Frankish king Childeric I at Tournai, Belg.

Attila is said to have brought the stirrup to Europe. Round or triangular iron stirrups were used by the Avars in the 6th century AD, and metal stirrups were used by the Byzantine cavalry. They were in use in China and Japan by about AD 600.

The principle of controlling a horse by exerting pressure on its mouth through a bit (a metal contrivance inserted in the mouth of the horse) and reins (straps attached to the bit held by the rider) was practiced from the earliest times, and bits made of bone and antlers have been found dating from before 1000 BC. The flexible mouthpiece with two links and its variations have been in use down the centuries, leading directly to the jointed snaffle bit of the present day.

Early, stumpy prickspurs have been found in Bohemia on 4th-century-BC Celtic sites.

Military horsemanship. The importance of cavalry increased in the early Middle Ages, and in the 1,000 years that followed, mounted warriors became predominant in battle. Armour steadily became bulkier and heavier, forcing the breeding of more and more massive horses, until the combination rendered maneuverability nearly impossible.

Efforts to overcome this were made at a Naples riding academy in the early 16th century, when Federico Grisone and Giovanni Battista Pignatelli tried to combine classical Greek principles with the requirements of medieval mounted combat. After Xenophon, except for a 14th-century treatise by Ibn Hudhayl, an Arab of Granada, Spain, apparently no literature on riding was produced until Grisone published his *Gli ordini di cavalcare* ("The Orders of Riding") in 1550.

The development of firearms led to the shedding of armour, making it possible for some further modifications in methods and training under followers of the school of Pignatelli and Grisone, such as William Cavendish, duke of Newcastle. In 1733 François Robichon de la Guérinière published *École de cavalerie* ("School of Cavalry"), in which he explained how a horse can be trained without being forced into submission, the fundamental precept of modern dressage. Dressage is the methodical training of a horse for any of a wide range of purposes, excluding only racing and cross-country riding.

Meanwhile, the Spanish Imperial Riding School in Vienna and the French cavalry centre at Saumur aimed at perfecting the combined performance of horse and rider. Their technique and academic seat, a formal riding position or style in which the rider sits erectly, deep in the middle of the saddle, exerted considerable influence in Europe and America during the 18th and 19th centuries and are still used in modern dressage. The head riding master at Saumur, Comte Antoine d'Aure, however, promoted a bold, relaxed, and more natural, if less "correct," style of riding across country, in disagreement with his 19th-century contemporary François Baucher, a horseman of great ability with formal haute école ("high school") ideas. Classical exercises in the manège, or school for riding, had to make way for simplified and more rational riding in war and the hunt. During this period hunting riders jumped obstacles with their feet forward, their torso back on the horse's haunches, and the horse's head held up. The horse often leaped in terror.

Growth of modern technique

At the turn of the 20th century, Capt. Federico Caprilli, an Italian cavalry instructor, made a thorough study of the psychology and mechanics of locomotion of the horse. He completely revolutionized the established system by innovating the forward seat, a position and style of riding in which the rider's weight is centred forward in the saddle, over the horse's withers. Caprilli wrote very little, but his pupil, Piero Santini, popularized his master's fundamental principles. Except in dressage and showing, the forward seat is the one now most frequently used, especially for jumping.

THE ART OF HORSEMANSHIP

The basic principle of horsemanship is to obtain results in a humane way by a combination of balance, seat, hands, and legs.

Fundamentals. The horse's natural centre of gravity shifts with its every movement and change of gait. Considering that a mounted horse also carries a comparatively unstable burden approximately one-fifth of its own weight, it is up to the rider to conform with the movements of the horse as much as possible.

Before one mounts, the saddle is checked to be sure that it fits both the horse and its rider. Experienced riders position themselves in the saddle in such a way as to be able to stay on the horse and control it. The seat adopted depends on the particular task at hand. A secure seat is essential, giving riders complete independence and freedom to apply effectively the aids at their disposal. Good riders do not overrule the horse, but, firmly and without inflicting pain, they persuade it to submit to their wishes.

The horse's movements. The natural gaits of the horse are the walk, the trot, the canter or slow gallop, and the gallop, although in dressage the canter and gallop are not usually differentiated. A riding horse is trained in each gait and in the change from one to another.

During the walk and the gallop the horse's head moves down and forward, then up and back (only at the trot is it still); riders follow these movements with their hands.

Walk. The walk is a slow, four-beat, rhythmic pace of distinct successive hoof beats in an order such as near (left) hind, near fore, off (right) hind, off fore. Alternately two or three feet may be touching the ground simultaneously. It may be a free, or ordinary, walk in which relaxed extended action allows the horse freedom of its head and neck, but contact with the mouth is maintained; or it may be a collected walk, a short-striding gait full of impulsion, or vigour; or it may be an extended walk of long, unhurried strides.

Trot. The trot is a two-beat gait, light and balanced, the fore and hind diagonal pairs of legs following each other almost simultaneously—near fore, off hind, off fore, and near hind. Riders can either sit in the saddle and be bumped as the horse springs from one diagonal to the other, or they can rise to the trot, post, by rising out of the saddle slightly and allowing more of their weight to bear on the stirrups when one or the other of the diagonal pairs of legs leaves the ground. Posting reduces the impact of the trot on both horse and rider.

Canter. As the horse moves faster, its gait changes into the canter, or ordinary gallop, in which the rider does not rise or bump. It is a three-beat gait, graceful and elegant, characterized by one or the other of the forelegs and both hindlegs leading—near hind, off hind, and near fore practically together, then off fore, followed briefly by complete suspension. Cantering can be on the near lead or the off, depending on which is the last foot to leave the ground. The rider's body is more forward than at the trot, the weight taken by the stirrups.

Gallop. An accelerated canter becomes the gallop, in which the rider's weight is brought sharply forward as the horse reaches speeds up to 30 miles (48 kilometres) an hour. The horse's movements are the same as in the canter. To some authorities, the gallop is a four-beat gait, especially in an extended run.

Other gaits. There are a number of disconnected and intermediate gaits, some done only by horses bred to perform them. One is the rack, a four-beat gait, with each beat evenly spaced in perfect cadence and rapid succession. The legs on either side move together, the hindleg striking the ground slightly before the foreleg. The single foot is similar to the rack but slower. In the pace, the legs on either side move and strike the ground together in a two-beat gait. The fox trot and the amble are four-beat gaits, the latter smoother and gliding.

The rack gait

Training and equipment. Depending on the abilities and inclinations of horse and trainer, training may include such elements as collection (controlled, precise, elevated movement) and extension (smooth, swift, reaching movement—the opposite of collection) at all paces; turns on the forehand (that part of the horse that is in front of the rider) and hindquarters; changing lead leg at the canter; change of speed; reining back, or moving backward; lateral movements; and finally the refinements of dressage, jumping, and cross-country riding.

Communication with the horse is rendered possible by the use of the bit and the aids. The rider signals intentions to the horse by a combination of recognized movements of hands and legs, using several articles of equipment. By repetition the horse remembers this language, understands what is required, and obeys.

Bits. There are several types of bits, including the snaffle, the double bridle, and the Pelham.

The simplest is the snaffle, also called the bridoon. It consists of a single straight or jointed mouthpiece with a ring at each end for the reins. The snaffle is used for racing and frequently for cross-country riding. It is appropriate for preliminary schooling.

The double bridle is used for advanced schooling. It consists of a jointed snaffle and a straight bit placed together in the mouth, first the snaffle, then the bit, both functioning independently and attached to separate reins. The mouthpiece of the bit can have a port or indentation in its centre to give more control. The slightest pull on the bit rein exerts pressure on the mouth.

The double bridle

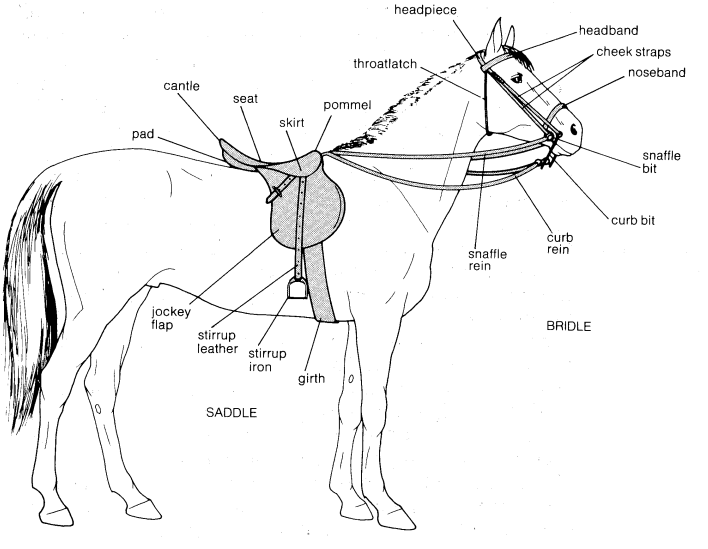
The Pelham is a snaffle with a straight mouthpiece; cheekpieces with rings at the lower ends for curb action; and a curb chain, with which pressure may be applied to the lower outside of the mouth. The Pelham gives control with only slight discomfort and is popular for polo.

Bridles. The bridle is a set of straps that makes the bit secure in the animal's mouth and thus ensures human control by means of the reins (see figure). The upper portion of the bridle consists of the headpiece passing behind the ears and joining the headband over the forehead; the cheek straps run down the sides of the head to the bit, to which they are fastened; in the blind type of driving bridle the blinkers, rectangular or round leather flaps that prevent the animal from seeing anything except what lies in front, are attached to the cheek straps; the noseband passes around the front of the nose just above the nostrils; and the throatlatch extends from the top of the cheek straps underneath the head.

Aids. The principal features of a horse's mentality are acute powers of observation, innate timidity, and a good memory. To a certain extent the horse can also understand. Schooling is based on these faculties, and the rider's aids are applied accordingly. The natural aids are the voice, the hands through the reins and the bit, the legs and heels, and the movement of the rider's weight. The whip, the spur, and devices such as martingales, special nosebands, and reins are artificial aids, so termed in theory, as the horse does not discriminate between natural and artificial.

Horses are easily startled. A good horseman will approach them quietly, speaking to them and patting them to give them confidence. Silence on the part of the rider can even cause disquiet to some horses, but they should not be shouted at. The rider's voice and its tone make

By courtesy of the American Saddle Horse Breeders Assoc. and W.W. Rouch & Co.



Nomenclature of a modern bridle and English saddle.

a useful aid in teaching a horse in its early schooling to walk, trot, canter, and halt.

To keep the horse alert at all times, the rider's hands keep a light, continual contact with its mouth, even at the halt. The hands are employed together with the legs to maintain contact, to urge the horse forward, to turn, to rein back, and generally to control the forehand. The horse is said to be collected and light in hand when the action of the bit can cause it to flex, or relax, its jaw with its head bent at the poll, or top.

When pressed simultaneously against the flanks, immediately after the hands ease the reins, the legs induce the forward movement of the horse. They are of the greatest importance in creating and maintaining impulsion, in controlling the hindquarters, and for lateral movement.

Riders achieve unity of balance by means of the weight aid, that is, by moving the body in harmony with the movements of the horse, forward, backward, or to the side. Thus, in cantering to the left, the rider leans to the left; or when about to descend a steep slope, the rider stays erect while the horse is feeling for the edge with its forefeet, but as soon as the descent starts the rider leans forward, leaving the hindquarters free to act as a brake and to prevent scraping the back of the horse's rear legs on rough ground. Meanwhile the hands keep the horse headed straight to maintain its balance.

The use of the whip

The whip is used chiefly to reinforce the leg aid for control, to command attention, and to demand obedience, but it can be used as a punishment in cases of deliberate rebellion. A horse may show resistance by gnashing its teeth and swishing its tail. Striking should always be on the quarters, behind the saddle girth, and must be immediate since a horse can associate only nearly simultaneous events. This applies equally to rewards. A friendly tone of voice or a pat on the neck are types of reward.

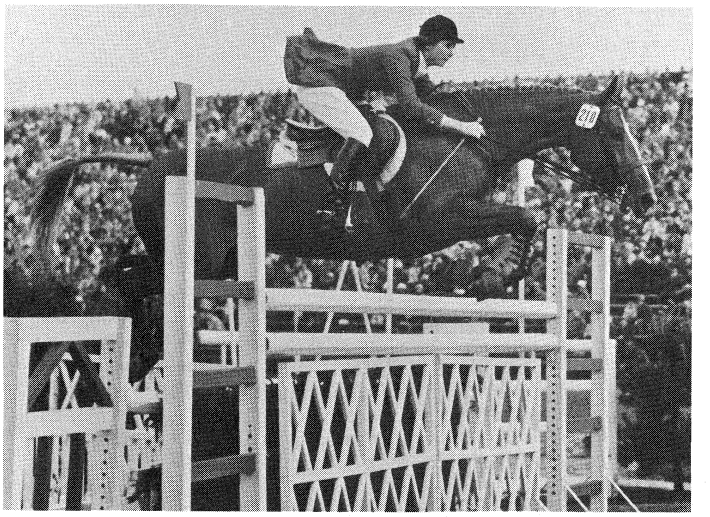
Although normally the leg or the heel, or both, should be sufficient, spurs, which should always be blunt, assist the legs in directing the precision movements of advanced schooling. Their use must be correctly timed.

Martingales are of three types: running, standing, or Irish. The running and standing martingales are attached to the saddle straps at one end and the bit reins or bridle at the other. The Irish martingale, a short strap below the horse's chin through which the reins pass, is used for racing and stops the horse from jerking the reins over its head. As the horse cannot see below a line from the eye to the nostril, it should not be allowed to toss its head back, particularly near an obstacle, as it is liable to leap blindly. A martingale should not be necessary with a well-schooled horse.

The noseband, a strap of the bridle that encircles the horse's nose, may be either a cavesson, with a headpiece and rings for attaching a long training rein, or a noseband with a headstrap, only necessary if a standing martingale is used. A variety of other nosebands are intended for horses that pull, or bear, on the reins unnecessarily.

Seats. The saddle, the length of the stirrup, and the rider's seat, or style of riding, should suit the purpose for which the horse is ridden. The first use of the stirrup is to enable the rider to get on the horse, normally from the near (left) side. With the raised foot in the stirrup the rider should avoid digging the horse in the flank on springing up and should gradually slide into position without landing on the horse's kidneys with a bump. With an excitable horse, the rider may wait, resting on knees and stirrups, until the horse moves forward.

Forward seat. The forward seat, favoured for show jumping, hunting, and cross-country riding, is generally considered to conform with the natural action of the horse. The rider sits near the middle of the saddle, his torso a trifle forward, even at the halt. The saddle is shaped with the flaps forward, sometimes with knee rolls for added support in jumping. The length of the stirrup leather is such that, with continual lower thigh and knee grip, the arch of the foot can press on the tread of the iron with the heel well down. A wide and heavy stirrup iron allows easy release of the foot in case of accidents. The line along the forearm from the elbow to the hands and along the reins to the bit is held straight. As the horse moves forward, so do the rider's hands, to suit the horse's comfort.

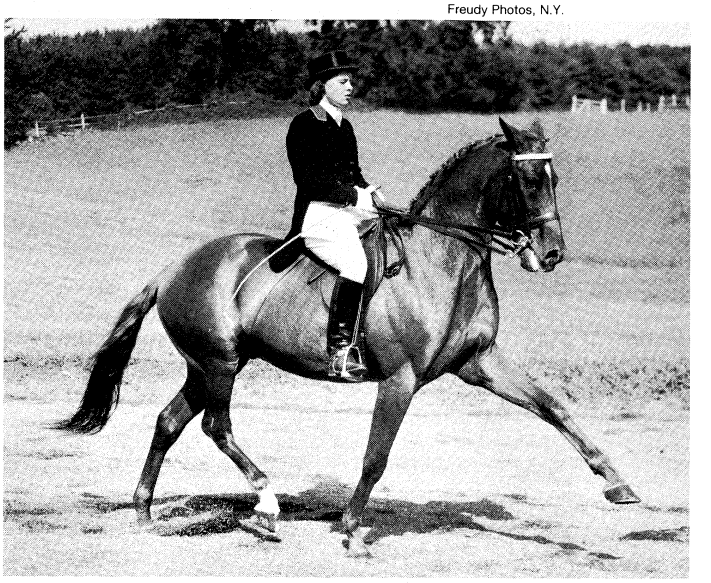


Forward seat in jumping.
By courtesy of the American Horse Shows Association

Dressage seat. In the show and dressage seat the rider sinks deep into the saddle, in a supple, relaxed but erect position above it. The saddle flaps are practically straight so as to show as much expanse of the horse's front as possible. The stirrup leather is of sufficient length for the rider's knee to bend at an angle of about 140 degrees and for the calf to make light contact with the horse's flank, the heel well down, and the toes or the ball of the foot resting on the tread of the stirrup iron. The rider keeps continual, light contact with the horse's mouth; and the intention is to convey an impression of graceful, collected action. In the past this type of saddle, with its straight-cut flaps, was used for hunting and polo, but the forward seat has become more popular for these activities.

Stock saddle. The stock saddle seat is appropriate for ranchers but is also used at rodeos and by many pleasure and trail riders. The saddle, which can weigh as much as 40 pounds (18 kilograms), is designed for rounding up cattle and is distinguished by a high pommel horn for tying a lariat. The rider employs long stirrups and a severe bit that he seldom uses because he rides with a loose rein, guiding his horse chiefly by means of shifting the weight of his body in the saddle. The gaucho roughriders of the Argentine Pampa have adopted a similar seat, using a saddle with a high pommel and cantle. Australian stockmen have used a saddle that has a short flap and is equipped with knee and thigh rolls, or props, which give an extremely secure seat.

Freudy Photos, N.Y.



Dressage seat in the extended trot.

Side saddle. Though now not so fashionable, the elegant and classical side-saddle seat was formerly favoured and considered correct by many horsewomen. On the near side the saddle has an upright pommel on which the rider's right leg rests. There is a lower, or leaping, pommel, against which the left leg can push upward when grip is required, and a single stirrup. Although the rider sits with both legs on one side of the saddle, forward action to suit the movement of the horse is feasible in cross-country riding.

Bareback. Bareback means riding without saddle or blanket, the rider sitting in the hollow of the horse's back and staying there chiefly by balance. It is an uncomfortable seat but less so at the walk and the slow canter. When suffering from saddle galls horses are sometimes ridden bareback for exercise.

Dressage. Originally intended for military use, dressage training was begun early in the 16th century. The international rules for dressage are based on the traditions and practice of the best riding schools in the world. The following is an extract from these rules of the Fédération Équestre Internationale:

> Object and general principles.
>
> The object of dressage is the harmonious development of the physique and the ability of the horse. As a result, it makes the horse calm, supple, and keen, thus achieving perfect understanding with its rider. These qualities are revealed by the freedom and regularity of the paces; the harmony, lightness, and ease of the movements; the lightening of the forehand, and the engagement of the hindquarters; the horse remaining absolutely straight in any movement along a straight line, and bending accordingly when moving on curved lines.
>
> The horse thus gives the impression of doing of his own account what is required of him. Confident and attentive, he submits generously to the control of his rider. (Used with permission of the publisher.)

Campagne is the term used for elementary but thorough training, including work on the longeing rein. This long rein, also used for training young or difficult horses, is attached to a headpiece with a noseband called a cavesson. The horse is bitted and saddled and is schooled in circles at the end of the rein. It is an accessory to training from the saddle, which is always best. Basic to campagne is collection: teaching the horse to arch its neck, to shift its weight backward onto its hindquarters, and to move in a showy, animated manner. Other elements of campagne include riding in a straight line, turns, and lateral movements.

Haute école dressage

Haute école is the most elaborate and specialized form of dressage, reaching its ultimate development at the Vienna school in its traditional white Lippizaner horses. Some characteristic haute école airs, or movements, are the pirouettes, which are turns on the haunches at the walk and the canter; the piaffe, in which the horse trots without moving forward, backward, or sideways, the impulse being upward; the passage, high-stepping trot in which the impulse is more upward than forward; the levade, in which the horse stands balanced on its hindlegs, its forelegs drawn in; the courvet, which is a jump forward in the levade position; and the croupade, ballotade, and capriole, a variety of spectacular airs in which the horse jumps and lands again in the same spot.

All of these movements are based, perhaps remotely in some instances, on those that the horse performs naturally.

Jumping. The most sensitive parts of the horse when ridden are the mouth and the loins, particularly in jumping. The rider's hands control the forehand while the legs act on the hindquarters. As speed is increased the seat is raised slightly from the saddle, with the back straight and the trunk and hands forward, the lower thighs and the knees taking the weight of the body and gripping the saddle, leaving the legs from the knees down free for impulsion. Contact with the mouth is maintained evenly and continually, the rider conforming with every movement as the horse's head goes forward after takeoff and as it is retracted on landing, the hands always moving in line with the horse's shoulder. In order to give complete freedom to the hindquarters and to the hocks, the rider does not sit back in the saddle until at least two strides after landing.

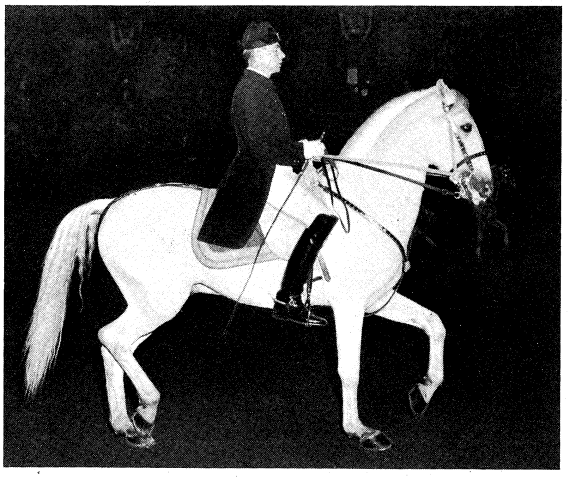
The horse is a natural jumper, but, if ridden, schooling becomes necessary. Training is started in an enclosed level area by walking the horse, preferably in a snaffle, over a number of bars or poles laid flat on the ground. When the horse has become accustomed to this, its speed is increased. As the horse progresses, the obstacles are systematically raised, varied, and spaced irregularly. The object is to teach the horse: (1) to keep its head down; (2) to approach an obstacle at a quiet, collected, yet energetic pace; (3) to decide how and where to take off; and (4) after landing to proceed quietly to the next obstacle. The horse should be confident over every jump before it is raised and should be familiarized with a variety of obstacles.

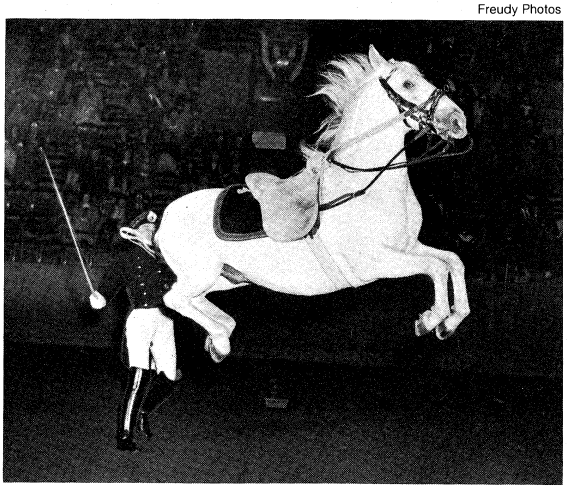
Only thoroughly trained riders and horses compete. Very strenuous effort is required of the horse, as well as of the rider who does not by any action give the horse the impression that something out of the ordinary is impending. If possible the horse is warmed up by at least a half-hour's walking and trotting before entering the ring. The horse is guided toward the exact centre of every obstacle, the rider looking straight ahead and not looking around after takeoff for any reason, as that might unbalance the horse. The broader the obstacle, the greater the speed of approach. Although a few experienced riders can adjust the horse's stride for a correct takeoff, this should not be necessary with a well-schooled horse. The rider is always made to conform with every action of the horse, the only assistance necessary being that of direction and increasing or decreasing speed according to the obstacle.

RIDING AND SHOWS

Racing on horseback probably originated soon after man first mastered the horse. By the 7th century BC organized mounted games were held at Olympia. The Romans held race meetings, and in medieval Europe tournaments, jousting, and horse fairs were frequent and popular events.

Freudy Photos





Haute école figures: Lippizaner horses in (left) piaffe and (right) ballotade.

Played in Persia for centuries, polo was brought to England from India about 1870. In North America, Western ranch riding produced the rodeo.

Horse associations and pony clubs are today the mainstay of equine sport. They have improved the standards of riding instruction and the competitive activities of dressage, hunter trials, and show jumping. The latter became an important event from 1869, when what was probably the first "competition for leaping horses" was included in the program of an Agricultural Hall Society horse show in London. National organizations such as the British Horse Society, the American Horse Shows Association (AHSA), the Federazione Italiana Sports Equestri, the National Equestrian Federation of Ireland, the Fédération Française des Sports Équestres, and similar groups from about 50 other nations are affiliated with the Fédération Équestre Internationale (FEI), founded in 1921 with headquarters at Brussels, the official international governing body and the authority on the requirements of equitation.

Associations and clubs

Horse shows. Horse shows are a popular institution that evolved from the horse sections of agricultural fairs. Originally they were informal displays intended to attract buyers and encourage the improvement of every type of horse. Now they are organized and conducted by committees of experts and by associations that enforce uniform rules, appoint judges, settle disputes, maintain records, and disseminate information. Riding contests included in the program have become increasingly important.

Under the auspices of the Royal Dublin Society, an international horse show was first held at Dublin in 1864. It is an annual exhibition of every type of saddle horse, as well as broodmares and ponies. International jumping contests similar to Olympic competition, events for children, and auction sales are held during this five-day show.

The National Horse Show at New York, first held in 1883, is another great yearly event. Held at Madison Square Garden, it lasts several days and includes about 10 different events. Among the most important are the international jumping under FEI rules and the open jumping under AHSA rules. Other shows are held in many sections of the United States.

Horse and pony shows are held regularly in the United Kingdom, the most important being the Richmond Royal Horse Show, the Horse of the Year Show, and the Royal International Horse Show. The latter, an annual event first held in 1907, has flourished under royal patronage and includes international jumping, special items such as the visit of the Spanish Riding School with its Lippizaners in 1953, and a Supreme Riding Horse competition.

In Canada, the Royal Agricultural Winter Fair at Toronto, opened in 1922 and known in Canada as the "Royal," is a major event, and in Australia the Royal Agricultural Society organizes horse shows annually in every state. Other events include the shows at Verona and at the Piazza di Siena in Rome; frequent horse shows in Belgium, France, Germany, and The Netherlands; the winter show in July in Buenos Aires; and the Exhibition of Economic Achievement in Moscow.

Olympic equestrian competition. The FEI organizes and controls the equestrian events at the Olympic Games. Included in each Olympics since the Games at Stockholm in 1912 (equestrian events were also held in 1900), these events are the occasion for keen rivalry and evoke high standards of horsemanship. They comprise a dressage grand prix, a three-day event, and a jumping grand prix, all open to team and individual competition.

The Grand Prix de Dressage involves performance of the walk, trot, canter, and collected paces and several conventional dressage figures and movements, as well as the correct rider's position. Scoring on each item is from a maximum of 10 for excellent down to 1 for very bad.

The three-day event consists of tests in dressage, endurance or cross-country riding, and show jumping. Dressage is on the first day. On the second day there is an endurance test over a course 25 to 35 kilometres (16 to 22 miles) in length, covering swamp roads, tracks, steeplechase obstacles, and cross-country sections. Jumping tests, less strenuous than the Prix des Nations jumping event, are held on the third day.

The Prix des Nations jumping event is a competition involving 13 or 14 obstacles, heights varying between 1.30 and 1.60 metres (51 and 63 inches), and a water jump 4 metres (13 feet) across, over a course with 60 metres (200 feet) between obstacles. Penalties are scored for disobedience, knocking down or touching an obstacle, and for a fall. The rider with the lowest penalty score wins.

In addition to these competitions there is a riding section of the modern pentathlon, also conducted under FEI rules. Competitors must clear, riding a strange horse chosen by lot, 20 obstacles over a course 1,000 metres (3,000 feet) in length. Other international competitions began in the 1950s under the supervision of the FEI. (C.E.C.)

BIBLIOGRAPHY

General works: GEORGE G. SIMPSON, *Horses* (1951, reprinted 1970), a very readable and popular account of the horse family today and through 60,000,000 years of development, with a good bibliography; MARGARET C. SELF, *The Horseman's Encyclopedia,* rev. ed. (1963, reprinted 1978), an invaluable collection of information on domestic horses. Breed associations issue pamphlets on selected breeds of horses. See also *A Standard Guide to Horse and Pony Breeds,* general ed. ELWYN HARTLEY EDWARDS (1980), providing information on 150 breeds and types. *The Horseman's International Book of Reference,* ed. by JEAN FROISSARD and LILY POWELL FROISSARD (1980), is an authoritative source consisting of contributions of international experts; C.E.G. HOPE and G.N. JACKSON (eds.), *The Encyclopedia of the Horse* (1973), is a comprehensive reference work discussing, among other specific topics, the horse in mythology, literature, and art. See also JOHN BASKETT, *The Horse in Art* (1980). Special aspects are examined in DAVID P. WILLOUGHBY, *Growth and Nutrition in the Horse* (1975); R.H. SMYTHE, *The Horse: Structure and Movement,* 2nd ed., rev. by PETER C. GOODY (1972); GEORGE H. WARING, *Horse Behavior: The Behavioral Traits and Adaptations of Domestic and Wild Horses, Including Ponies* (1983); MOYRA WILLIAMS, *Horse Psychology,* rev. ed. (1976); and PETER WILLETT, *The Thoroughbred* (1970). R.S. SUMMERHAYS, *Encyclopaedia for Horsemen,* 6th ed., rev. by STELLA A. WALKER (1975), is also a useful reference work.

Horsemanship: MARY GORDON-WATSON, *The Handbook of Riding* (1982); ALBERT E. DECARPENTRY, *Academic Equitation,* trans. from the French (1971); VLADIMIR S. LITTAUER, *Common Sense Horsemanship,* 2nd ed. (1963, reprinted 1974); HENRY WYNMALEN, *Equitation,* 2nd ed. (1952, reissued 1971); EARL R. FARSCHLER, *Riding and Training,* new ed. (1959, reissued 1972), which contains a description of the gaits; JEAN S.-F. PAILLARD, *Understanding Equitation,* trans. from the French (1974); C.E.G. HOPE, *The Horseman's Manual* (1972); JANE KIDD, *Horsemanship in Europe* (1977); and MYRON J. SMITH, *Equestrian Studies* (1981), a classified bibliography of more than 4,600 English-language items appearing between 1950 and 1980. See also JACKIE SPAULDING, *The Family Horse: How to Choose, Care for, Train and Work Your Horse* (1982). (*History*): STAN STEINER, *Dark and Dashing Horsemen* (1981); CHARLES CHENEVIX TRENCH, *A History of Horsemanship* (1970); and GLENN R. VERNAM, *Man on Horseback* (1964, reissued 1972), which includes information on the origin and detail of equipment. (*Horse shows*): R.S. SUMMERHAYS and C.E.G. HOPE, *Horse Shows: The Judges, Stewards, Organizers* (1969); *American Horse Shows Association Rule Book* (biennial); HARLAN C. ABBEY, *Horses and Horse Shows* (1980); JUDY RICHTER, *Horse and Rider: From Basics to Show Competition* (1982); EDWARD HART, *Care and Showing of the Heavy Horse* (1981); and LYNDA BLOOM, *Fitting and Showing the Halter Horse* (1980). (*Rules*): The rules for international competitions are given in publications of the FÉDÉRATION ÉQUESTRE INTERNATIONALE; in BOB PHILLIPS (ed.), *Official Report of the Olympic Games* (1969); and in various publications of the BRITISH HORSE SOCIETY. (*Dressage*): RICHARD L. WÄTJEN, *Dressage Riding: A Guide for the Training of Horse and Rider,* 3rd rev. ed. (1979; originally published in German, 1922; 7th German ed., 1975); HANS HANDLER, *The Spanish Riding School,* trans. from the German (1972); ELWYN HARTLEY EDWARDS, *Saddlery: Modern Equipment for Horse and Stable* (1963, reissued 1973); and *The USCTA Book of Eventing: The Official Handbook of the United States Combined Training Association,* ed. by SALLY O'CONNOR (1982). (*Jumping*): J.A. TALBOT-PONSONBY, *Harmony in Horsemanship* (1964, reissued 1972); F.C. AVIS, *Horses and Show Jumping Dictionary,* 2nd ed., ed. by MARGARET B. SLESSOR (1979); and FEDERICO CAPRILLI, *The Caprilli Papers: Principles of Outdoor Equitation,* trans. and ed. by P. SANTINI (1967).

(A.W.P./C.E.C./Ed.)

Human Rights

It is a common observation that human beings everywhere demand the realization of diverse values to ensure their individual and collective well-being. It also is a common observation that these demands are often painfully frustrated by social as well as natural forces, resulting in exploitation, oppression, persecution, and other forms of deprivation. Deeply rooted in these twin observations are the beginnings of what today are called "human rights" and the legal processes, national and international, associated with them.

This article is divided into the following sections:

Historical development 656
Definition of human rights 657
  Nature 658
  Content 658
  Legitimacy and priority 659
International human rights: prescription and enforcement 660
  Before World War II 660
  Human rights in the United Nations 660
  Human rights and the Helsinki Process 662
  Regional developments 662
  International human rights in domestic courts 663
Conclusion 664
Bibliography 664

Historical development

The expression "human rights" is relatively new, having come into everyday parlance only since World War II and the founding of the United Nations in 1945. It replaces the phrase "natural rights," which fell into disfavour in part because the concept of natural law (to which it was intimately linked) had become a matter of great controversy, and the later phrase "the rights of Man," which was not universally understood to include the rights of women.

Origins

Most students of human rights trace the historical origins of the concept back to ancient Greece and Rome, where it was closely tied to the premodern natural law doctrines of Greek Stoicism (the school of philosophy founded by Zeno of Citium, which held that a universal working force pervades all creation and that human conduct therefore should be judged according to, and brought into harmony with, the law of nature). The classic example, drawn from the Greek literature, is that of Antigone, who, upon being reproached by Creon for defying his command not to bury her slain brother, asserted that she acted in accordance with the immutable laws of the gods.

In part because Hellenistic Stoicism played a key role in its formation and spread, Roman law may similarly be seen to have allowed for the existence of a natural law and, with it, pursuant to the *jus gentium* ("law of nations"), certain universal rights that extended beyond the rights of citizenship. According to the Roman jurist Ulpian, for example, natural law was that which nature—not the state—assures to all human beings, Roman citizen or not.

It was not until after the Middle Ages, however, that natural law doctrines became closely associated with liberal political theories about natural rights. In Greco-Roman and medieval times, natural law doctrines taught mainly the duties, as distinguished from the rights, of "Man." Moreover, as evident in the writings of Aristotle and St. Thomas Aquinas, these doctrines recognized the legitimacy of slavery and serfdom and, in so doing, excluded perhaps the centralmost ideas of human rights as they are understood today—the ideas of freedom (or liberty) and equality.

For the idea of human (*i.e.*, natural) rights to take hold as a general social need and reality, it was necessary that basic changes in the beliefs and practices of society take place, changes of the sort that evolved from about the 13th century to the Peace of Westphalia (1648), during the Renaissance and the decline of feudalism. When resistance to religious intolerance and political–economic bondage began the long transition to liberal notions of freedom and equality, particularly in relation to the use and ownership of property, then were the foundations of what today are called human rights truly laid. During this period, reflecting the failure of rulers to meet their natural law obligations as well as the unprecedented commitment to individual expression and worldly experience that was characteristic of the Renaissance, the shift from natural law as duties to natural law as rights was made. The teachings of Aquinas (1224/25–1274) and Hugo Grotius (1583–1645) on the European continent, and the Magna Carta (1215), the Petition of Right of 1628, and the English Bill of Rights (1689) in England, were proof of this change. All testified to the increasingly popular view that human beings are endowed with eternal and inalienable rights, never renounced when humankind "contracted" to enter the social from the primitive state and never diminished by the claim of "the divine right of kings."

Influence of the Renaissance

It was primarily for the 17th and 18th centuries, however, to elaborate upon this modernist conception of natural law as meaning or implying natural rights. The scientific and intellectual achievements of the 17th century—the discoveries of Galileo and Sir Isaac Newton, the materialism of Thomas Hobbes, the rationalism of René Descartes and Gottfried Wilhelm Leibniz, the pantheism of Benedict de Spinoza, the empiricism of Francis Bacon and John Locke—encouraged a belief in natural law and universal order; and during the 18th century, the so-called Age of Enlightenment, a growing confidence in human reason and in the perfectability of human affairs led to its more comprehensive expression. Particularly to be noted are the writings of the 17th-century English philosopher John Locke—arguably the most important natural law theorist of modern times—and the works of the 18th-century Philosophes centred mainly in Paris, including Montesquieu, Voltaire, and Jean-Jacques Rousseau. Locke argued in detail, mainly in writings associated with the Revolution of 1688 (the Glorious Revolution), that certain rights self-evidently pertain to individuals as human beings (because they existed in "the state of nature" before humankind entered civil society); that chief among them are the rights to life, liberty (freedom from arbitrary rule), and property; that, upon entering civil society (pursuant to a "social contract"), humankind surrendered to the state only the right to enforce these natural rights, not the rights themselves; and that the state's failure to secure these reserved natural rights (the state itself being under contract to safeguard the interests of its members) gives rise to a right to responsible, popular revolution. The Philosophes, building on Locke and others and embracing many and varied currents of thought with a common supreme faith in reason, vigorously attacked religious and scientific dogmatism, intolerance, censorship, and social–economic restraints. They sought to discover and act upon universally valid principles harmoniously governing nature, humanity, and society, including the theory of the inalienable "rights of Man" that became their fundamental ethical and social gospel.

Ideas of Locke

All this liberal intellectual ferment had, not surprisingly, great influence on the Western world of the late 18th and early 19th centuries. Together with the practical example of England's Revolution of 1688 and the resulting Bill of Rights, it provided the rationale for the wave of revolutionary agitation that then swept the West, most notably in North America and France. Thomas Jefferson, who had studied Locke and Montesquieu and who asserted

that his countrymen were a "free people claiming their rights as derived from the laws of nature and not as the gift of their Chief Magistrate," gave poetic eloquence to the plain prose of the 17th century in the Declaration of Independence proclaimed by the 13 American Colonies on July 4, 1776: "We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the Pursuit of Happiness." Similarly, the Marquis de Lafayette, who won the close friendship of George Washington and who shared the hardships of the American War of Independence, imitated the pronouncements of the English and American revolutions in the Declaration of the Rights of Man and of the Citizen of August 26, 1789. Insisting that "men are born and remain free and equal in rights," the declaration proclaims that "the aim of every political association is the preservation of the natural and imprescriptible rights of man," identifies these rights as "Liberty, Property, Safety and Resistance to Oppression," and defines "liberty" so as to include the right to free speech, freedom of association, religious freedom, and freedom from arbitrary arrest and confinement (as if anticipating the Bill of Rights added in 1791 to the Constitution of the United States of 1787).

In sum, the idea of human rights, called by another name, played a key role in the late 18th- and early 19th-century struggles against political absolutism. It was, indeed, the failure of rulers to respect the principles of freedom and equality, which had been central to natural law philosophy almost from the beginning, that was responsible for this development. In the words of Maurice Cranston, a leading student of human rights, "... absolutism prompted man to claim [human, or natural] rights precisely because it denied them."

The idea of human rights as natural rights was not without its detractors, however, even at this otherwise receptive time. In the first place, being frequently associated with religious orthodoxy, the doctrine of natural rights became less and less acceptable to philosophical and political liberals. Additionally, because they were conceived in essentially absolutist—"inalienable," "unalterable," "eternal"—terms, natural rights were found increasingly to come into conflict with one another. Most importantly, the doctrine of natural rights came under powerful philosophical and political attack from both the right and the left.

Burke, Hume, and Bentham

In England, for example, conservatives Edmund Burke and David Hume united with liberal Jeremy Bentham in condemning the doctrine, the former out of fear that public affirmation of natural rights would lead to social upheaval, the latter out of concern lest declarations and proclamations of natural rights substitute for effective legislation. In his *Reflections on the Revolution in France* (1790), Burke, a believer in natural law who nonetheless denied that the "rights of Man" could be derived from it, criticized the drafters of the Declaration of the Rights of Man and of the Citizen for proclaiming the "monstrous fiction" of human equality, which, he argued, serves but to inspire "false ideas and vain expectations in men destined to travel in the obscure walk of laborious life." Bentham, one of the founders of Utilitarianism and a nonbeliever, was no less scornful. "Rights," he wrote, "is the child of law; from real law come real rights; but from imaginary laws, from 'law of nature,' come imaginary rights.... Natural rights is simple nonsense; natural and imprescriptible rights (an American phrase), rhetorical nonsense, nonsense upon stilts." Hume agreed with Bentham; natural law and natural rights, he insisted, are unreal metaphysical phenomena.

This assault upon natural law and natural rights, thus begun during the late 18th century, both intensified and broadened during the 19th and early 20th centuries. John Stuart Mill, despite his vigorous defense of liberty, proclaimed that rights ultimately are founded on utility. The German jurist Friedrich Karl von Savigny, England's Sir Henry Maine, and other historicalists emphasized that rights are a function of cultural and environmental variables unique to particular communities. And the jurist John Austin and the philosopher Ludwig Wittgenstein insisted, respectively, that the only law is "the command of the sovereign" (a phrase of Thomas Hobbes) and that the only truth is that which can be established by verifiable experience. By World War I, there were scarcely any theorists who would or could defend the "rights of Man" along the lines of natural law. Indeed, under the influence of 19th-century German Idealism and parallel expressions of rising European nationalism, there were some—the Marxists, for example—who, although not rejecting individual rights altogether, maintained that rights, from whatever source derived, belong to communities or whole societies and nations preeminently. Thus did F.H. Bradley, the British Idealist, write in 1894: "The rights of the individual are today not worth serious consideration.... The welfare of the community is the end and is the ultimate standard."

Yet, though the heyday of natural rights proved short, the idea of human rights nonetheless endured in one form or another. The abolition of slavery, factory legislation, popular education, trade unionism, the universal suffrage movement—these and other examples of 19th-century reformist impulse afford ample evidence that the idea was not to be extinguished even if its transempirical derivation had become a matter of general skepticism. But it was not until the rise and fall of Nazi Germany that the idea of rights—human rights—came truly into its own. The laws authorizing the dispossession and extermination of Jews and other minorities, the laws permitting arbitrary police search and seizure, the laws condoning imprisonment, torture, and execution without public trial—these and similar obscenities brought home the realization that law and morality, if they are to be deserving of the name, cannot be grounded in any purely Utilitarian, Idealist, or other consequentialist doctrine. Certain actions are wrong, no matter what; human beings are entitled to simple respect at least.

Human rights in the 20th century

Today, the vast majority of legal scholars, philosophers, and moralists agree, irrespective of culture or civilization, that every human being is entitled, at least in theory, to some basic rights. Heir to the Protestant Reformation and to the English, American, French, Mexican, Russian, and Chinese revolutions, the last half of the 20th century has seen, in the words of human rights scholar Louis Henkin, "essentially universal acceptance of human rights in principle" such that "no government dares to dissent from the ideology of human rights today." Indeed, except for some essentially isolated 19th-century demonstrations of international humanitarian concern to be noted below, the last half of the 20th century may fairly be said to mark the birth of the international as well as the universal recognition of human rights. In the treaty establishing the United Nations (UN), all members pledged themselves to take joint and separate action for the achievement of "universal respect for, and observance of, human rights and fundamental freedoms for all without distinction as to race, sex, language, or religion." In the Universal Declaration of Human Rights (1948), representatives from many diverse cultures endorsed the rights therein set forth "as a common standard of achievement for all peoples and all nations." And in 1976, the International Covenant on Economic, Social and Cultural Rights and the International Covenant on Civil and Political Rights, each approved by the UN General Assembly in 1966, entered into force and effect.

Definition of human rights

To say that there is widespread acceptance of the principle of human rights on the domestic and international planes is not to say that there is complete agreement about the nature of such rights or their substantive scope—which is to say, their definition. Some of the most basic questions have yet to receive conclusive answers. Whether human rights are to be viewed as divine, moral, or legal entitlements; whether they are to be validated by intuition, custom, social contract theory, principles of distributive justice, or as prerequisites for happiness; whether they are to be understood as irrevocable or partially revocable; whether they are to be broad or limited in number and content—these and kindred issues are matters of ongoing debate and likely will remain so as long as there ex-

ist contending approaches to public order and scarcities among resources.

NATURE

Despite this lack of consensus, however, a number of widely accepted—and interrelated—postulates may be seen to assist, if not to complete, the task of defining human rights. Five in particular stand out, although it is to be noted that not even these are without controversy.

First, regardless of their ultimate origin or justification, human rights are understood to represent individual and group demands for the shaping and sharing of power, wealth, enlightenment, and other cherished values in community process, most fundamentally the value of respect and its constituent elements of reciprocal tolerance and mutual forebearance in the pursuit of all other values. Consequently, they imply claims against persons and institutions who impede realization and standards for judging the legitimacy of laws and traditions. At bottom, human rights limit state power.

Second, reflecting varying environmental circumstances, differing worldviews, and inescapable interdependencies within and between value processes, human rights refer to a wide continuum of value claims ranging from the most justiciable to the most aspirational. Human rights partake of both the legal and the moral orders, sometimes indistinguishably. They are expressive of both the "is" and the "ought" in human affairs.

Universal nature of human rights

Third, if a right is determined to be a human right it is quintessentially general or universal in character, in some sense equally possessed by all human beings everywhere, including in certain instances even the unborn. In stark contrast to "the divine right of kings" and other such conceptions of privilege, human rights extend, in theory, to every person on Earth without discriminations irrelevant to merit.

Fourth, most assertions of human rights—arguably not all—are qualified by the limitation that the rights of any particular individual or group in any particular instance are restricted as much as is necessary to secure the comparable rights of others and the aggregate common interest. Given this interdependency, human rights are sometimes designated prima facie rights, and it makes little or no sense to think or talk of them in absolutist terms.

Fifth and finally, human rights are commonly assumed to refer, in some vague sense, to "fundamental" as distinct from "nonessential" claims or "goods." In fact, some theorists go so far as to limit human rights to a single core right or two—for example, the right to life or the right to equal freedom of opportunity. The tendency, in short, is to de-emphasize or rule out "mere wants."

In several critical respects, however, this last postulate raises more questions than it answers. What does it mean to say that a right is fundamental? Does it entail some bare minimum only, or, more plausibly, does it admit to something greater? If the latter, how much greater and subject to what conditions, if any? In other words, however accurate, this last postulate is fraught with ambiguity about the content and legitimate scope of human rights and about the priorities, if any, that obtain among them. Except for the issue of the origin and justification of human rights, no cluster of preliminary human rights considerations is more controversial.

CONTENT

It cannot be disputed that, like all normative traditions, the human rights tradition is a product of its time. It necessarily reflects the processes of historical continuity and change that, at once and as a matter of cumulative experience, help to give it substance and form. Therefore, to understand better the debate over the content and legitimate scope of human rights and the priorities claimed among them, it is useful to note the dominant schools of thought and action that have informed the human rights tradition since the beginning of modern times.

Particularly helpful in this regard is the notion of "three generations of human rights" advanced by the French jurist Karel Vasak. Inspired by the three normative themes of the French Revolution, they are: the first generation of civil and political rights (*liberté*); the second generation of economic, social, and cultural rights (*égalité*); and the third generation of newly called solidarity rights (*fraternité*). Vasak's model is of course a simplified expression of an extremely complex historical record; it is not intended as a literal representation of life in which one generation gives birth to the next and then dies away.

Civil and political rights

The first generation. The first generation of civil and political rights derives primarily from the 17th- and 18th-century reformist theories noted above, which are associated with the English, American, and French revolutions. Infused with the political philosophy of liberal individualism and the economic and social doctrine of laissez-faire, it conceives of human rights more in negative ("freedoms from") than positive ("rights to") terms; it favours the abstention rather than the intervention of government in the quest for human dignity, as epitomized by the statement attributed to H.L. Mencken that " . . . all government is, of course, against liberty." Belonging to this first generation, thus, are such claimed rights as are set forth in Articles 2–21 of the Universal Declaration of Human Rights, including freedom from racial and equivalent forms of discrimination; the right to life, liberty, and the security of the person; freedom from slavery or involuntary servitude; freedom from torture and from cruel, inhuman, or degrading treatment or punishment; freedom from arbitrary arrest, detention, or exile; the right to a fair and public trial; freedom from interference in privacy and correspondence; freedom of movement and residence; the right to asylum from persecution; freedom of thought, conscience, and religion; freedom of opinion and expression; freedom of peaceful assembly and association; and the right to participate in government, directly or through free elections. Also included is the right to own property and the right not to be deprived of one's property arbitrarily, each fundamental to the interests fought for in the American and French revolutions and to the rise of capitalism.

Of course, it would be error to assert that these and other first-generation rights correspond completely to the idea of "negative" as opposed to "positive" rights. The right to security of the person, to a fair and public trial, to asylum from persecution, and to free elections, for example, manifestly cannot be assured without some affirmative government action. What is constant in this first-generation conception, however, is the notion of liberty, a shield that safeguards the individual, alone and in association with others, against the abuse and misuse of political authority. This is the core value. Featured in almost every constitution of today's approximately 160 states, and dominating the majority of the international declarations and covenants adopted since World War II, this essentially Western liberal conception of human rights is sometimes romanticized as a triumph of Hobbesian–Lockean individualism over Hegelian statism.

Economic, social, and cultural rights

The second generation. The second generation of economic, social, and cultural rights finds its origins primarily in the socialist tradition that was foreshadowed among the Saint-Simonians of early 19th-century France and variously promoted by revolutionary struggles and welfare movements ever since. In large part, it is a response to the abuses and misuses of capitalist development and its underlying, essentially uncritical, conception of individual liberty that tolerated, even legitimated, the exploitation of working classes and colonial peoples. Historically, it is counterpoint to the first generation of civil and political rights, with human rights conceived more in positive ("rights to") than negative ("freedoms from") terms, requiring the intervention, not the abstention, of the state for the purpose of assuring equitable participation in the production and distribution of the values involved. Illustrative are the claimed rights set forth in Articles 22–27 of the Universal Declaration of Human Rights, such as the right to social security; the right to work and to protection against unemployment; the right to rest and leisure, including periodic holidays with pay; the right to a standard of living adequate for the health and well-being of self and family; the right to education; and the right to the protection of one's scientific, literary, and artistic production.

Yet, in the same way that all the rights embraced by the first generation of civil and political rights cannot properly be designated "negative rights," so all the rights embraced by the second generation of economic, social, and cultural rights cannot properly be labeled "positive rights." The right to free choice of employment, the right to form and to join trade unions, and the right freely to participate in the cultural life of the community, for example, do not inherently require affirmative state action to ensure their enjoyment. Nevertheless, most of the second-generation rights do necessitate state intervention in the allocation of resources because they subsume demands more for material than for intangible values according to some criterion of distributive justice. Second-generation rights are, fundamentally, claims to social equality. Partly because of the comparatively late arrival of socialist–communist influence in the normative domain of international affairs, however, the internationalization of these rights has been somewhat slow in coming; but with the ascendancy of the Third World on the global stage, intent upon a "revolution of rising expectations," the second-generation rights have begun to come of age.

Solidarity rights

The third generation. Finally, the third generation of solidarity rights, while drawing upon, interlinking, and reconceptualizing value demands associated with the two earlier generations of rights, are best understood as a product, albeit one still in formation, of both the rise and the decline of the nation-state in the last half of the 20th century. Foreshadowed in Article 28 of the Universal Declaration of Human Rights, which proclaims that "everyone is entitled to a social and international order in which the rights set forth in this Declaration can be fully realized," it appears so far to embrace six claimed rights. Three of these reflect the emergence of Third World nationalism and its demand for a global redistribution of power, wealth, and other important values: the right to political, economic, social, and cultural self-determination; the right to economic and social development; and the right to participate in and benefit from "the common heritage of mankind" (shared Earth–space resources; scientific, technical, and other information and progress; and cultural traditions, sites, and monuments). The other three third-generation rights—the right to peace, the right to a healthy and balanced environment, and the right to humanitarian disaster relief—suggest the impotence or inefficiency of the nation-state in certain critical respects.

All six of these claimed rights tend to be posed as collective rights, requiring the concerted efforts of all social forces, to substantial degree on a planetary scale, and implying a quest for a possible utopia that projects the notion of holistic community interests. Each, however, manifests an individual as well as collective dimension. For example, while it may be said to be the collective right of all countries and peoples (especially developing countries and non-self-governing peoples) to secure a new international economic order that would eliminate obstacles to their economic and social development, so also may it be said to be the individual right of all persons to benefit from a developmental policy that is based on the satisfaction of material and nonmaterial human needs. Also, while the right to self-determination and the right to humanitarian assistance, for example, find expression on the legal as well as the moral plane, the majority of these solidarity rights tend to be more aspirational than justiciable in character, enjoying as yet an ambiguous jural status as international human rights norms.

Thus, at various stages of modern history—following the "bourgeois" revolutions of the 17th and 18th centuries, the socialist and Marxist revolutions of the early 20th century, and the anticolonialist revolutions that began immediately following World War II—the content of human rights has been broadly defined, not with any expectation that the rights associated with one generation would or should become outdated upon the ascendancy of another, but expansively or supplementally. Reflecting evolving perceptions of which values, at different times, stand most in need of encouragement and protection, the history of the content of human rights also reflects humankind's recurring demands for continuity and stability.

LEGITIMACY AND PRIORITY

This is not to imply that each of these three generations of rights is equally acceptable to all or that they or their separate elements are greeted with equal urgency. First-generation proponents, for example, are inclined to exclude second- and third-generation rights from their definition of human rights altogether (or, at best, to label them as "derivative"). In part this is due to the complexities that inform the process of putting these rights into action. The suggestion of greater feasibility that attends first-generation rights because they stress the absence rather than the presence of government is somehow transformed into a prerequisite of a comprehensive definition of human rights, such that aspirational and vaguely asserted claims to entitlement are deemed not to be rights at all. The most forceful explanation, however, is more ideologically or politically motivated. Persuaded that egalitarian claims against the rich, particularly where collectively espoused, are unworkable without a severe decline in liberty and quality (in part because they involve state intervention for the redistribution of privately held resources), first-generation proponents, inspired by the natural law and laissez-faire traditions, are partial to the view that human rights are inherently independent of civil society and are individualistic.

Conversely, second- and third-generation defenders often look upon first-generation rights, at least as commonly practiced, as insufficiently attentive to material human needs and, indeed, as legitimating instruments in service to unjust domestic, transnational, and international social orders—hence constituting a "bourgeois illusion." Accordingly, while not placing first-generation rights outside their definition of human rights, they tend to assign such rights a low status and therefore to treat them as long-term goals that will come to pass only with fundamental economic and social transformations to be realized progressively and fully consummated only sometime in the future.

In sum, different conceptions of rights, particularly emerging conceptions, contain the potential for challenging the legitimacy and supremacy not only of one another but, more importantly, of the political–social systems with which they are most intimately associated. As a consequence there is sharp disagreement about the legitimate scope of human rights and about the priorities that are claimed among them.

A misleading debate

On final analysis, however, this liberty–equality and individualist–collectivist debate over the legitimacy and priorities of claimed human rights can be dangerously misleading. It is useful, certainly, insofar as it calls attention to the way in which notions of liberty and individualism can be, and have been, used to rationalize the abuses of capitalism; and it is useful, too, insofar as it highlights how notions of equality and collectivism can be, and have been, alibis for authoritarian governance. But in the end it risks obscuring at least three essential truths that must be taken into account if the contemporary worldwide human rights movement is to be objectively understood.

First, one-sided characterizations of legitimacy and priority are likely, over the long term, to undermine the political credibility of their proponents and the defensibility of their particularistic values. In an increasingly interdependent and interpenetrating global community, any human rights orientation that does not genuinely support the widest possible shaping and sharing of all values among all human beings is likely to provoke widespread skepticism. The last half of the 20th century is replete with examples.

Second, such characterizations do not accurately mirror behavioral reality. In the real world, despite differences in cultural tradition and ideological style, there exists a rising and overriding insistence upon the equitable production and distribution of all basic values. U.S. Pres. Franklin D. Roosevelt's Four Freedoms (freedom of speech and expression, freedom of worship, freedom from want, and freedom from fear) is an early case in point. A more recent demonstration was the 1977 Law Day speech by then U.S. Secretary of State Cyrus R. Vance, in which he announced the U.S. government's resolve "to make the advancement of human rights a central part of our foreign policy" and defined human rights to include "the right to

be free from governmental violation of the integrity of the person, . . . the right to the fulfillment of such vital needs as food, shelter, health care, and education, . . . [and] the right to enjoy civil and political liberties." Essentially individualistic societies tolerate, even promote, certain collectivist values; likewise, essentially communal societies tolerate, even promote, certain individualistic values. Ours is a more-or-less, not an either-or, world.

Finally, none of the international human rights instruments currently in force or proposed say anything whatsoever about the legitimacy or rank-ordering of the rights they address, save possibly in the case of rights that by international covenant are stipulated to be nonderogable and therefore, arguably, more fundamental than others (for example, freedom from arbitrary or unlawful deprivation of life, freedom from torture and from inhuman or degrading treatment and punishment, freedom from slavery, freedom from imprisonment for debt). There is disagreement, to be sure, among lawyers, moralists, and political scientists about the legitimacy and hierarchy of claimed rights when they treat the problem of implementation. For example, some insist on certain civil and political guarantees, whereas others defer initially to conditions of material and corporeal well-being. Such disagreements, however, partake of political agendas and have little if any conceptual utility. As the UN General Assembly has repeatedly confirmed, all human rights form an indivisible whole.

In short, the legitimacy of different human rights and the priorities claimed among them are a function of context. Because people in different parts of the world both assert and honour different human rights demands according to many different procedures and practices, these issues ultimately depend on time, place, setting, level of crisis, and other circumstance.

International human rights: prescription and enforcement

BEFORE WORLD WAR II

International Law of State Responsibility for Injuries to Aliens

Ever since ancient times, but especially since the emergence of the modern state system, the Age of Discovery, and the accompanying spread of industrialization and European culture throughout the world, there has developed, for economic and other reasons, a unique set of customs and conventions relative to the humane treatment of foreigners. This evolving International Law of State Responsibility for Injuries to Aliens, as these customs and conventions came to be called, may be understood to represent the beginning of active concern for human rights on the international plane. The founding fathers of international law—particularly Francisco de Vitoria (1486?–1546), Hugo Grotius (1583–1645), and Emmerich de Vattel (1714–67)—were quick to observe that all persons, outlander as well as other, were entitled to certain natural rights; and they emphasized, consequently, the importance of according aliens fair treatment.

Except, however, for the occasional use of treaties to secure the protection of Christian minorities, as early illustrated by the Peace of Westphalia (1648), which concluded the Thirty Years' War and established the principle of equal rights for the Roman Catholic and Protestant religions in Germany, it was not until the start of the 19th century that active international concern for the rights of nationals began to make itself felt. Then, in the century and a half before World War II, several noteworthy, if essentially unconnected, efforts to encourage respect for nationals by international means began to shape what today is called the International Law of Human Rights (which for historical but no theoretically convincing reasons has tended to be treated separately from the International Law of State Responsibility for Injuries to Aliens).

Throughout the 19th and early 20th centuries, numerous military operations and diplomatic representations, not all of them with the purest of motives but done nonetheless in the name of "humanitarian intervention" (a customary international law doctrine), undertook to protect oppressed and persecuted minorities in the Ottoman Empire and in Syria, Crete, various Balkan countries, Romania, and Russia. Paralleling these actions, first at the Congress of Vienna (1814–15) and later between the two world wars, a series of treaties and international declarations sought the protection of certain racial, religious, and linguistic minorities in central and eastern Europe and in the Middle East. During the same period the movement to combat and suppress slavery and the slave trade found expression in treaties sooner or later involving the major commercial powers, beginning with the Treaty of Paris (1814) and culminating in the International Slavery Convention (1926).

In addition, toward the end of the 19th century and continuing well beyond World War II, the community of nations, inspired largely by persons associated with what is now the International Committee of the Red Cross, concluded a series of multilateral declarations and agreements designed to temper the conduct of hostilities, protect the victims of war, and otherwise elaborate the humanitarian law of war. At about the same time, first with two multilateral labour conventions concluded in 1906 and subsequently at the initiative of the International Labour Organisation (ILO; established in 1919), a reformist-minded international community embarked upon a variety of collaborative measures directed at the promotion of human rights. These included not only fields traditionally associated with labour law and relations (for example, industrial health, safety, and welfare; hours of work; annual paid holidays) but also—mainly after World War II—in respect of such core human rights concerns as forced labour, discrimination in employment and occupation, freedom of association for collective bargaining, and equal pay for equal work.

Finally, during the interwar period, the Covenant establishing the League of Nations (1919), while not formally recognizing "the rights of Man" and while failing to lay down a principle of racial nondiscrimination as requested by Japan (owing mainly to the resistance of Great Britain and the United States), nevertheless committed the League's members to several human rights goals: fair and humane working conditions for men, women, and children; the execution of agreements regarding traffic in women and children; the prevention and control of disease in matters of international concern; and the just treatment of native colonial peoples. Also, victorious powers who as "mandatories" were entrusted by the League with the tutelage of colonies formerly governed by Germany and Turkey accepted as "a sacred trust of civilization" responsibilities for the well-being and development of the inhabitants of those territories. (The arrangement was carried over into the UN trusteeship system and had serious repercussions more than a half century later in relation to the mandate entrusted to South Africa over the territory of South West Africa [now Namibia].)

Human rights goals of the League of Nations

As important as these pre-World War II human rights efforts were, however, it was not until after the War—and the Nazi atrocities accompanying it—that active concern for human rights on the international plane truly came of age. In the proceedings of the International Military Tribunal at Nürnberg in 1945–46, German high officials were tried not only for "crimes against peace" and "war crimes" but also for "crimes against humanity" committed against any civilian population even if in accordance with the laws of the country where perpetrated. While the tribunal, whose establishment and rulings subsequently were endorsed by the UN General Assembly, applied a cautious approach to allegations of "crimes against humanity," it nonetheless made the treatment by a state of its own citizens the subject of international criminal process.

HUMAN RIGHTS IN THE UNITED NATIONS

The Charter of the United Nations (1945) begins by reaffirming a "faith in fundamental human rights, in the dignity and worth of the human person, in the equal rights of men and women and of nations large and small." It states that the purposes of the United Nations are, among other things, "to develop friendly relations among nations based on respect for the principle of equal rights and self-determination of peoples . . . [and] to achieve international co-operation . . . in promoting and encouraging respect for human rights and for fundamental freedoms for all with-

out distinction as to race, sex, language, or religion . . . ." And, in two key articles, all members "pledge themselves to take joint and separate action in cooperation with the Organization" for the achievement of these and related purposes. It is to be noted, however, that a proposal to ensure the protection as well as the promotion of human rights was explicitly rejected at the San Francisco Conference establishing the United Nations. Additionally, the Charter expressly provides that nothing in it "shall authorize the United Nations to intervene in matters which are essentially within the domestic jurisdiction of any state . . . ," except upon a Security Council finding of a "threat to the peace, breach of the peace, or act of aggression." Moreover, although typical of major constitutive instruments, the Charter is conspicuously general and vague in its human rights clauses, among others.

Thus, not surprisingly, the reconciliation of the Charter's human rights provisions with the Charter's drafting history and its "domestic jurisdiction" clause has given rise to not a little legal and political controversy. Some authorities have argued that, in becoming parties to the Charter, states accept no more than a nebulous promotional obligation toward human rights and that, in any event, the United Nations has no standing to insist on human rights safeguards in member states. Others insist that the Charter's human rights provisions, being part of a legally binding treaty, clearly involve some element of legal obligation; that the "pledge" made by states upon becoming party to the Charter consequently represents more than a moral statement; and that the "domestic jurisdiction" clause does not apply because human rights, whatever isolation they may have "enjoyed" in the past, no longer can be considered matters "essentially within the domestic jurisdiction" of states.

When all is said and done, however, it is clear from the actual practice of the United Nations that the problem of resolving these opposing contentions has proved somewhat less formidable than the statements of governments and the opinions of scholars might lead one to assume. Neither the Charter's drafting history nor its "domestic jurisdiction" clause nor, indeed, its generality and vagueness in respect of human rights has prevented the United Nations—on the basis of individual petitions, statements from witnesses, state complaints, and reports from interested nongovernmental organizations—from investigating, discussing, and evaluating specific human rights situations. Nor have they prevented it from recommending or prescribing concrete action in relation to them, at least not in the case of "a consistent pattern of gross violations" of human rights, provided there has been a majority persuasive enough to force the action desired (as in the imposition by the Security Council in 1977 of a mandatory arms embargo against South Africa). Of course, governments usually are protective of their sovereignty (or domestic jurisdiction). Also, the UN organs responsible for the promotion of human rights suffer from most of the same disabilities that afflict the United Nations as a whole, in particular the absence of supranational authority and the presence of divisive power politics. Hence, it cannot be expected that UN actions in defense of human rights will be, normally, either swift or categorically effective. Nevertheless, assuming some political will, the legal obstacles to UN enforcement of human rights are not insurmountable.

UN Commission on Human Rights

Primary responsibility for the promotion of human rights under the UN Charter rests in the General Assembly and, under its authority, in the Economic and Social Council and its subsidiary body, the Commission on Human Rights, an intergovernmental body that serves as the UN's central policy organ in the human rights field. Much of the commission's activity, initiated by subsidiary working groups, is investigatory, evaluative, and advisory in character, and the commission annually establishes a working group to consider and make recommendations concerning alleged "gross violations" of human rights referred to it by its Sub-Commission on Prevention of Discrimination and Protection of Minorities (on the basis of communications from individuals and groups, pursuant to Resolution 1503 [1970] of the UN Economic and Social Council, and sometimes on the basis of investigations by the subcommission or one of its working groups). Also, the commission has appointed special representatives and envoys to examine human rights situations on an ad hoc basis, who, in the course of preparing their reports, examine reliable information submitted in good faith, interview interested persons, or make on-site inspections with the cooperation of the government concerned.

In addition, the commission, together with other UN organs such as the International Labour Organisation (ILO), the UN Educational, Scientific and Cultural Organization (UNESCO), and the UN Commission on the Status of Women, drafts human rights standards and has prepared a number of international human rights instruments. Among the most important are the Universal Declaration of Human Rights (1948), the International Covenant on Civil and Political Rights (together with its Optional Protocol; 1976), and the International Covenant on Economic, Social and Cultural Rights (1976). Collectively known as the International Bill of Rights, these three instruments serve as touchstones for interpreting the human rights provisions of the UN Charter.

The Universal Declaration of Human Rights. The catalog of rights set out in the Universal Declaration of Human Rights, which was adopted without dissent by the General Assembly on December 10, 1948, is scarcely less than the sum of all the important traditional political and civil rights of national constitutions and legal systems, including equality before the law; protection against arbitrary arrest; the right to a fair trial; freedom from ex post facto criminal laws; the right to own property; freedom of thought, conscience, and religion; freedom of opinion and expression; and freedom of peaceful assembly and association. Also enumerated are such economic, social, and cultural rights as the right to work and to choose one's work freely, the right to equal pay for equal work, the right to form and join trade unions, the right to rest and leisure, the right to an adequate standard of living, and the right to education.

The Universal Declaration, it must be noted, is not a treaty. It was meant to proclaim "a common standard of achievement for all peoples and all nations" rather than enforceable legal obligations. Nevertheless, partly because of an 18-year delay between its adoption and the completion for signature and ratification of the two covenants, the Universal Declaration has acquired a status juridically more important than originally intended. It has been widely used, even by national courts, as a means of judging compliance with human rights obligations under the UN Charter.

The International Covenant on Civil and Political Rights and the Optional Protocol. The civil and political rights guaranteed by this covenant, which was opened for signature on December 19, 1966, and entered into force on March 23, 1976, incorporate almost all of those proclaimed in the Universal Declaration, including the right to nondiscrimination. Pursuant to the covenant, each state party undertakes to respect and to ensure to all individuals within its territory and subject to its jurisdiction the rights recognized in the covenant "without distinction of any kind, such as race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status." Some rights listed in the Universal Declaration, however, such as the right to own property and the right to asylum, are not included among the rights recognized in the covenant. Similarly, the covenant designates a number of rights that are not listed in the Universal Declaration, among them the right of all peoples to self-determination and the right of ethnic, religious, or linguistic minorities to enjoy their own culture, to profess and practice their own religion, and to use their own language. To the extent that the Universal Declaration and the covenant overlap, however, the latter is understood to explicate and help interpret the former.

In addition, the covenant calls for the establishment of a Human Rights Committee, an international organ of 18 persons elected by the parties to the covenant, serving in their individual expert capacity and charged to study reports submitted by the state parties on the measures they have adopted that give effect to the rights recognized

in the covenant. As between the state parties that have expressly recognized the competence of the committee in this regard, the committee also may respond to allegations by one state party that another state party is not fulfilling its obligations under the covenant. If the committee is unable to resolve the problem, the matter is referred to an ad hoc conciliation commission, which eventually reports its findings on all questions of fact, plus its views on the possibilities of an amicable solution. State parties that become party to the Optional Protocol further recognize the competence of the Human Rights Committee similarly to consider and act upon communications from individuals claiming to be victims of covenant violations.

The International Covenant on Economic, Social and Cultural Rights. Just as the International Covenant on Civil and Political Rights elaborates upon most of the civil and political rights enumerated in the Universal Declaration of Human Rights, so the International Covenant on Economic, Social and Cultural Rights elaborates upon most of the economic, social, and cultural rights set forth in the Universal Declaration: the right to work; the right to just and favourable conditions of work; trade union rights; the right to social security; rights relating to the protection of the family; the right to an adequate standard of living; the right to health; the right to education; and rights relating to culture and science. Unlike its companion International Covenant on Civil and Political Rights, however, this covenant is not geared, with modest exception, to immediate implementation, the state parties having agreed only "to take steps" toward "achieving progressively the full realization of the rights recognized in the . . . Covenant," and then subject to "the maximum of [their] available resources." The covenant is essentially a "promotional convention," stipulating objectives more than standards and requiring implementation over time rather than all at once. One obligation is, however, subject to immediate application: the prohibition of discrimination in the enjoyment of the rights enumerated on grounds of race, colour, sex, language, religion, or political or other opinion; national or social origin; property; and birth or other status. Also, the international supervisory measures that apply to the covenant oblige the state parties to report to the UN Economic and Social Council on the steps they have adopted and the progress they have made in achieving the realization of the enumerated rights.

Other UN human rights conventions. The two abovementioned covenants are by no means the only human rights treaties drafted and adopted under the auspices of the United Nations. Indeed, because there are far too many to detail even in abbreviated fashion, it must suffice simply to note that they address a broad range of concerns, including the prevention and punishment of the crime of genocide; the humane treatment of military and civilian personnel in time of war; the status of refugees; the protection and reduction of stateless persons; the abolition of slavery, forced labour, and discrimination in employment and occupation; the elimination of all forms of racial discrimination and the suppression and punishment of the crime of apartheid; the elimination of discrimination in education; the promotion of the political rights of women and the elimination of all forms of discrimination against women; and the promotion of equality of opportunity and treatment of migrant workers. (For particular agreements, see *Human Rights: A Compilation of International Instruments*, 3rd ed. [1978], published by the United Nations.) Many of these treaties are the work of the UN specialized agencies, particularly the International Labour Organisation (ILO), and many also provide for supervisory and enforcement mechanisms—for example, the Committee on the Elimination of Racial Discrimination established under the International Convention on the Elimination of All Forms of Racial Discrimination of December 21, 1965.

UN human rights declarations. In addition to developing human rights standards and procedures through treaties, the UN General Assembly, impressed by the impact of the Universal Declaration of Human Rights, also has resorted to the proclamation of declarations as a means of promoting human rights. Adopted in the form of a resolution of the General Assembly, which technically is not binding on the member states in the sense of a treaty, a declaration, particularly when it enunciates principles of great and solemn importance, may nevertheless create within the international community strong expectations about authority and control. Perhaps the best known examples subsequent to the Universal Declaration, while not devoted exclusively to human rights considerations, are the Declaration on the Granting of Independence to Colonial Countries and Peoples (1960) and the Declaration on Principles of International Law Concerning Friendly Relations and Co-Operation Among States in Accordance with the Charter of the United Nations (1970).

HUMAN RIGHTS AND THE HELSINKI PROCESS

Post-World War II concern for human rights also has been evident at the global level outside the United Nations, most notably in the proceedings and aftermath of the Conference on Security and Cooperation in Europe, convened in Helsinki on July 3, 1973, and concluded there (after continuing deliberations in Geneva) on August 1, 1975. Attended by representatives of 35 governments that included the NATO countries, the Warsaw Pact nations, and 13 neutral and nonaligned European states, the conference had as its principal purpose a mutually satisfactory definition of peace and stability between East and West, previously made impossible by the period of the Cold War. In particular, the Soviet Union was concerned with achieving recognition of its western frontiers as established at the end of World War II.

There was little tangible, however, that the Western powers, with no realistic territorial claims of their own, could demand in return, and accordingly they pressed for certain concessions in respect of human rights and freedom of movement and information between East and West. Thus, at the outset of the Final Act adopted by the conference, in a Declaration of Principles Guiding Relations Between States, the participating governments solemnly declared "their determination to respect and put into practice," alongside other "guiding" principles, "respect [for] human rights and fundamental freedoms, including the freedom of thought, conscience, religion or belief" and "respect [for] the equal rights of peoples and their right to self-determination." It was hoped that this would mark the beginning of a liberalization of authoritarian regimes.

The Helsinki Final Act

From the earliest discussions, however, it was clear that the Helsinki Final Act was not intended as a legally binding instrument. "Determination to respect" and "put into practice" were deemed to express moral commitments only, the Declaration of Principles was said not to prescribe international law, and nowhere did the participants provide for enforcement machinery. On the other hand, the Declaration of Principles, including its human rights principles, always has been viewed as at least consistent with international law. Additionally, the fourth of four sections (commonly known as "baskets") of the Final Act provides for the holding of periodic review conferences in which the participating states are called upon "to continue the multilateral process initiated by the Conference." But most importantly, ever since their adoption, the Final Act's human rights provisions have served as important and widely accepted yardsticks for external scrutiny and appropriate recourse to perceived violations.

In sum, like the Universal Declaration of Human Rights and other such declarations of the UN General Assembly, the Helsinki Final Act, though not a treaty, has created widespread expectations about proper human rights behaviour, and consequently it has inspired and facilitated the monitoring of human rights policy. Assuming some cordiality between East and West, the Helsinki Process may be said at least to hold out the potential for modestly beneficial results in the human rights arena.

REGIONAL DEVELOPMENTS

Action for the international promotion and protection of human rights has proceeded at the regional level in Europe, the Americas, Africa, and the Middle East. Only the first three of these regions, however, have gone so far as to create enforcement mechanisms within the framework of a human rights charter. The Permanent Arab Commission

on Human Rights, founded by the Council of the League of Arab States in September 1968 but since then preoccupied by the rights of Arabs living in Israeli-occupied territories, has not brought a proposed Arab Convention on Human Rights to a successful conclusion and so far has tended to function more in terms of the promotion than the protection of human rights.

European human rights system. On November 4, 1950, the Council of Europe agreed to the European Convention for the Protection of Human Rights and Fundamental Freedoms, the substantive provisions of which are based on a draft of what is now the International Covenant on Civil and Political Rights. Together with its five additional protocols, this convention, which entered into force on September 3, 1953, represents the most advanced and successful international experiment in the field. A companion instrument, similar to the later International Covenant on Economic, Social and Cultural Rights, is the European Social Charter (1961). The charter's provisions are implemented through an elaborate system of control based on the sending of progress reports to, and the appraisal of these reports by, the various committees and organs of the Council of Europe. The instrumentalities created under the European convention are the European Commission of Human Rights and the European Court of Human Rights. The convention also makes use of the governmental organ of the Council of Europe, the Committee of Ministers.

European Social Charter

The commission may receive from any state party to the convention any allegation of a breach of the convention by another state party. Also, provided its legal competence to do so has been formally recognized, the commission may receive petitions from any person, group of individuals, or nongovernmental organization claiming to be the victim of a violation of the convention. In such cases, the commission is charged to ascertain the facts and to place itself at the disposal of the parties to secure "a friendly settlement . . . on the basis of respect for Human Rights." If no such solution is reached, the commission is called upon to draw up a report, stating its opinion as to whether the facts disclose a breach, and to recommend action to the Committee of Ministers, including referral of the case to the European Court of Human Rights.

The jurisdiction of the court extends to cases referred to it by a state party whose national is alleged to be a victim of a violation, by a state party against whom a complaint has been lodged, and by any state party that may have referred the case to the commission. The court may not, however, receive a complaint by an individual applicant. Moreover, it may receive state complaints only if the defendant state has accepted its jurisdiction. This may be done ad hoc for a particular case or by a general declaration accepting the compulsory jurisdiction of the court. In either event, and in cases referred by the European commission as well, the judgment of the court is final. If a question is not or cannot be referred to the court, then the Committee of Ministers of the Council of Europe makes a final decision on human rights complaints.

The instrumentalities of the European convention have, over the years, developed a considerable body of case law on questions regulated by the convention; and the provisions of the convention are deemed, in some European states, part of domestic constitutional or statutory law. In countries where this is not the case, the state parties to the convention have taken other measures to make their domestic laws conform with their obligations under the convention.

Inter-American human rights system. In 1948, concurrent with its establishment of the Organization of American States (OAS), the Ninth Pan-American Conference adopted the American Declaration on the Rights and Duties of Man, an instrument similar to, but coming a full seven months before, the Universal Declaration of the United Nations and setting out the duties as well as the rights of the individual citizen (a throwback, perhaps, to Greco-Roman and medieval natural law theories). Subsequently, in 1959, a meeting of consultation of the American Ministers for Foreign Affairs created, within the framework of the OAS, the Inter-American Commission on Human Rights, which has since undertaken important investigative activities concerning human rights in the Americas. Finally, in 1969, the Inter-American Specialized Conference on Human Rights, meeting in San José, Costa Rica, adopted the American Convention on Human Rights, which made the existing Inter-American Commission on Human Rights an organ for the convention's implementation and established the Inter-American Court of Human Rights, which sits in San José.

Both the substantive law and the procedural arrangements of the American convention, which entered into force in 1978, are strongly influenced by the UN covenants and the European convention, and they were drafted also with the European Social Charter in mind. Under the American convention, however, unlike its UN and European predecessors, the right of petition by individuals, groups of individuals, and nongovernmental organizations operates automatically. Under the UN system, the right of petition applies only when the state concerned has become a party to the Optional Protocol to the International Covenant on Civil and Political Rights, and under the European system a special declaration by the states concerned is required. On the other hand, again in contrast to the European system (but not the UN system), interstate complaints under the American convention operate only among states that have expressly agreed to such procedure.

African human rights system. In 1981, following numerous pleas by the UN Commission on Human Rights, interested states, nongovernmental organizations, and others dating as far back as 1961, the Eighteenth Assembly of Heads of State and Government of the Organization of African Unity (OAU), convening in Nairobi, Kenya, adopted the African Charter on Human and Peoples' Rights. The charter became effective on October 21, 1986, after it was ratified by a majority of the 50 member states of the OAU.

The African Charter

Like its European and American counterparts, the African charter provides for the establishment of an African Commission on Human and Peoples' Rights, with both promotional and protective functions and with no restriction on who may file a complaint with the commission (thus signatory states, individuals, groups of individuals, and nongovernmental organizations, whether or not they are victims of the alleged violation, may all file). In contrast to the European and American procedures, however, concerned states are encouraged to reach a friendly settlement without formally involving the investigative or conciliatory mechanisms of the commission. Also, the African charter does not call for a human rights court. African customs and traditions, it is said, emphasize mediation, conciliation, and consensus rather than the adversarial and adjudicative procedures that are common to Western legal systems.

Four other distinctive features of the African charter are especially noteworthy. First, it provides for economic, social, and cultural rights as well as civil and political rights. In this respect it bears resemblance to the American convention, but it is distinctive from the European convention. Next, in contrast to both the European and American conventions, it recognizes the rights of groups in addition to the family, women, and children. The aged and the infirm are accorded special protection also, and the right of peoples to self-determination is elaborated in the right to existence, equality, and nondomination. Third, it uniquely embraces two third-generation, or "solidarity," rights "as belonging to all peoples": the right to economic, social, and cultural development and the right to national and international peace and security. Finally, it is so far the only treaty instrument to detail individual duties as well as individual rights—to the family, society, the state, and the international African community.

INTERNATIONAL HUMAN RIGHTS IN DOMESTIC COURTS

Using domestic courts to clarify and safeguard international human rights is a new and still evolving approach to human rights advocacy. In addition to the inevitable interpretative problems of applying conventional and customary norms that are fashioned in multicultural settings, controversial theories about the interrelation of national

and international law plus many procedural difficulties—carrying such labels as "standing," "act of State," and the "political questions doctrine"—burden the party anxious to invoke international human rights norms in the domestic context. To be sure, considerable progress has been made, as perhaps best evidenced in the far-reaching decision handed down by the U.S. Court of Appeals for the 2nd Circuit in 1980 in *Filartiga* v. *Pena-Irala*, in which the court held that the international prohibition of torture is unequivocally established in the law of nations and therefore to be honoured in U.S. courts. But as human rights scholar Richard Lillich has cautioned, " . . . in all likelihood the [national] judiciary will have to experience much more international human rights law consciousness-raising before [wholesale resistance to its domestic application] is rejected."

Conclusion

Whatever the current attitudes and policies of governments, the reality of popular demands for human rights, including both greater economic justice and greater political freedom, is beyond debate. A deepening and widening concern for the promotion and protection of human rights, hastened by the self-determinist impulse of a post-colonial era, is now unmistakably woven into the fabric of contemporary world affairs.

Substantially responsible for this progressive development has been, of course, the work of the United Nations, its allied agencies, and such regional organizations as the Council of Europe, the Organization of American States, and the Organization of African Unity. Also visibly helpful, however, particularly since the early 1970s, have been three other factors: the public advocacy of human rights as a key aspect of national foreign policies, made initially legitimate by the example of U.S. Pres. Jimmy Carter; the emergence and proliferation of activist nongovernmental human rights organizations such as Amnesty International (winner of the Nobel Prize for Peace for 1977), the International Commission of Jurists, and diverse church-affiliated groups; and a worldwide profusion of courses and materials devoted to the study of human rights both in formal and informal educational settings. Indeed, in light of the weaknesses that presently inhere at the intergovernmental level of global and regional organization, it is likely that each of these factors will play an increasingly important role in the future.

To be sure, formidable obstacles attend the endeavours of human rights policymakers, activists, and scholars. The implementation of international human rights law depends for the most part on the voluntary consent of nations; the mechanisms for the observance or enforcement of human rights are yet in their infancy. Still, it is certain that a palpable concern for the advancement of human rights is here to stay, out of necessity no less than out of idealism. As Nobel laureate and political dissident Andrey Sakharov once wrote from his internal exile in the Soviet Union:

> The ideology of human rights is probably the only one which can be combined with such diverse ideologies as communism, social democracy, religion, technocracy and those ideologies which may be described as national and indigenous. It can also serve as a foothold for those . . . who have tired of the abundance of ideologies, none of which have brought . . . simple human happiness. The defense of human rights is a clear path toward the unification of people in our turbulent world, and a path toward the relief of suffering.

BIBLIOGRAPHY

Documentary: UNITED NATIONS, *Human Rights: A Compilation of International Instruments,* 3rd ed. (1978), contains the texts of human rights treaties and other instruments established under the auspices of the United Nations. UNITED NATIONS, *Yearbook on Human Rights* (annual), documents national and international developments in the human rights field. See also JAMES AVERY JOYCE, *Human Rights: International Documents,* 3 vol. (1978); BURNS H. WESTON, RICHARD A. FALK, and ANTHONY A. D'AMATO (eds.), *Basic Documents in International Law and World Order* (1980); IAN BROWNLIE (comp.), *Basic Documents on Human Rights,* 2nd ed. (1981); RICHARD B. LILLICH (ed.), *International Human Rights Instruments: A Compilation of Treaties, Agreements, and Declarations of Especial Interest to the United States* (1983); and UNIFO, *International Human Rights Instruments of the United Nations* (1983).

General: Basic works on the subject include SIR HERSCH LAUTERPACHT, *International Law and Human Rights* (1950, reprinted 1973); JÓZSEF HÁLASZ (ed.), *Socialist Concept of Human Rights,* trans. from the Hungarian (1966); EGON SCHWELB, *Human Rights and the International Community: The Roots and Growth of the Universal Declaration of Human Rights* (1964); EVAN LUARD (ed.), *The International Protection of Human Rights* (1967); ASBJÖRN EIDE and AUGUST SCHOU (eds.), *International Protection of Human Rights: Proceedings of the Seventh Nobel Symposium, September 25–27, 1967* (1968); R. BILDER, "Rethinking International Human Rights: Some Basic Questions," *Wisconsin Law Review,* pp. 171–217, no. 1 (1969); JOHN CAREY, *UN Protection of Civil and Political Rights* (1970); VERNON VAN DYKE, *Human Rights, the United States, and World Community* (1970); MAURICE CRANSTON, *What Are Human Rights?* (1973); J. HUMPHREY, "The International Law of Human Rights in the Middle Twentieth Century," in MAARTEN BOS (ed.), *The Present State of International Law and Other Essays* (1973); MOSES MOSKOWITZ, *International Concern with Human Rights* (1974); MANOUCHEHR GANJI, *The Realization of Economic, Social, and Cultural Rights: Problems, Policies, Progress* (1975); RICHARD P. CLAUDE (ed.), *Comparative Human Rights* (1976); THOMAS BUERGENTHAL (ed.), *Human Rights, International Law, and the Helsinki Accord* (1977); FRANCISZEK PRZETACZNIK, "The Socialist Concept of Human Rights: Its Philosophical Background and Political Justification," *Belgian Review of International Law,* 13:239–278 (1977); FOUAD AJAMI, *Human Rights and World Order Politics* (1978); JAMES AVERY JOYCE, *The New Politics of Human Rights* (1979); B.G. RAMCHARAN (ed.), *Human Rights: Thirty Years After the Universal Declaration* (1979); MYRES S. MCDOUGAL, HAROLD D. LASSWELL, and LUNG-CHU CHEN, *Human Rights and World Public Order: The Basic Policies of an International Law of Human Dignity* (1980); RICHARD A. FALK, *Human Rights and State Sovereignty* (1981); LOUIS HENKIN (ed.), *The International Bill of Rights: The Covenant on Civil and Political Rights* (1981), and *The Rights of Man Today* (1978); S.P. MARKS, "Emerging Human Rights: A New Generation for the 1980s," *Rutgers Law Review,* 33:435–452 (Winter 1981); A.H. ROBERTSON, *Human Rights in the World: An Introduction to the Study of the International Protection of Human Rights,* 2nd ed. (1982); KAREL VASAK (ed.), *The International Dimension of Human Rights,* 2 vol., trans. from the French (1982); PAUL SIEGHART, *The International Law of Human Rights* (1983); and HURST HANNUM (ed.), *Guide to International Human Rights Practice* (1984).

Regional instruments and arrangements: J.E.S. FAWCETT, *The Application of the European Convention on Human Rights* (1969); S. MARKS, "La Commission permanente arabe des droits de l'homme," *Revue de droits de l'homme/Human Rights Journal,* 3:101–108 (1970); FREDE CASTBERG, *The European Convention on Human Rights,* trans. from the Norwegian (1974); COUNCIL OF EUROPE, *European Convention on Human Rights: Collected Texts* (1975); A.H. ROBERTSON, *Human Rights in Europe,* 2nd ed. (1977); THOMAS BUERGENTHAL and ROBERT E. NORRIS (eds.), *Human Rights: The Inter-American System* (1982–); R. GITTLEMAN, "The African Charter on Human and People's Rights: A Legal Analysis," *Virginia Journal of International Law,* 22:667–714 (1982); and U. UMOZURIKE, "The African Charter on Human and People's Rights," *American Journal of International Law,* 77:902–912 (October 1983).

Human rights journals: *Human Rights Law Journal* (quarterly); *Human Rights Quarterly* (formerly *Universal Human Rights*); *The Human Rights Review* (three times a year, 1976–81); and *Revue des droits de l'homme/Human Rights Journal* (quarterly, 1968–79).

(Bu.H.W.)

Humanism

The word *humanism* has been freely applied to a variety of beliefs, methods, and philosophies that place central emphasis on the human realm. Most frequently, however, the term is used with reference to a system of education and mode of inquiry that developed in northern Italy during the 14th century and later spread through Europe and England. Alternately known as "Renaissance humanism," this program was so broadly and profoundly influential that it is one of the chief reasons why the Renaissance is viewed as a distinct historical period. Indeed, though the word *Renaissance* is of more recent coinage, the fundamental idea of that period as one of renewal and reawakening is humanistic in origin. But humanism sought its own philosophical bases in far earlier times and, moreover, continued to exert some of its power long after the end of the Renaissance.

This article is divided into the following sections:

Origin and meaning of the term *humanism* 665
- The ideal of *humanitas* 665
- Other uses 665

Basic principles and attitudes 666
- Classicism 666
- Realism 666
- Critical scrutiny and concern with detail 666
- The emergence of the individual and the idea of the dignity of man 666
- Active virtue 667

Early history 667
The 15th century 669
- Leon Battista Alberti 669
- The Medici and Federico da Montefeltro 669

Later Italian humanism 670
- Things and words 670
- Idealism and the Platonic Academy of Florence 670
- Machiavelli's realism 671
- The achievement of Castiglione 671
- Tasso's Aristotelianism 671

Northern humanism 671
- Desiderius Erasmus 671
- The French humanists 672
- The English humanists 672

Humanism and the visual arts 674
- Realism 674
- Classicism 674
- Anthropocentricity and individualism 674
- Art as philosophy 674

Humanism, art, and science 675
Humanism and Christianity 675
Later fortunes of humanism 676
Conclusion 676
Bibliography 676

ORIGIN AND MEANING OF THE TERM HUMANISM

The ideal of humanitas. The history of the term *humanism* is complex but enlightening. It was first employed (as *humanismus*) by 19th-century German scholars to designate the Renaissance emphasis on classical studies in education. These studies were pursued and endorsed by educators known, as early as the late 15th century, as *umanisti:* that is, professors or students of classical literature. The word *umanisti* derives from the *studia humanitatis,* a course of classical studies that, in the early 15th century, consisted of grammar, poetry, rhetoric, history, and moral philosophy. The *studia humanitatis* were held to be the equivalent of the Greek *paideia.* Their name was itself based on the Latin *humanitas,* an educational and political ideal that was the intellectual basis of the entire movement. Renaissance humanism in all its forms defined itself in its straining toward this ideal. No discussion, therefore, of humanism can have validity without an understanding of *humanitas.*

The *studia humanitatis*

Humanitas meant the development of human virtue, in all its forms, to its fullest extent. The term thus implied not only such qualities as are associated with the modern word *humanity*—understanding, benevolence, compassion, mercy—but also such more aggressive characteristics as fortitude, judgment, prudence, eloquence, and even love of honour. Consequently the possessor of *humanitas* could not be merely a sedentary and isolated philosopher or man of letters but was of necessity a participant in active life. Just as action without insight was held to be aimless and barbaric, insight without action was rejected as barren and imperfect. *Humanitas* called for a fine balance of action and contemplation, a balance born not of compromise but of complementarity. The goal of such fulfilled and balanced virtue was political in the broadest sense of the word. The purview of Renaissance humanism included not only the education of the young but also the guidance of adults (including rulers) via philosophical poetry and strategic rhetoric. It included not only realistic social criticism but also utopian hypotheses, not only painstaking reassessments of history but also bold reshapings of the future. In short, humanism called for the comprehensive reform of culture, the transfiguration of what humanists termed the passive and ignorant society of the "dark" ages into a new order that would reflect and encourage the grandest human potentialities. Humanism had an evangelical dimension. It sought to project *humanitas* from the individual into the state at large.

The role of the classics

The wellspring of *humanitas* was classical literature. Greek and Roman thought, available in a flood of rediscovered or newly translated manuscripts, provided humanism with much of its basic structure and method. For Renaissance humanists, there was nothing dated or outworn about the writings of Plato, Cicero, or Livy. Compared with the typical productions of medieval Christianity, these pagan works had a fresh, radical, almost avant-garde tonality. Indeed, recovering the classics was to humanism tantamount to recovering reality. Classical philosophy, rhetoric, and history were seen as models of proper method—efforts to come to terms, systematically and without preconceptions of any kind, with perceived experience. Moreover, classical thought considered ethics qua ethics, politics qua politics: it lacked the inhibiting dualism occasioned in medieval thought by the often conflicting demands of secularism and Christian spirituality. Classical virtue, in examples of which the literature abounded, was not an abstract essence but a quality that could be tested in the forum or on the battlefield. Finally, classical literature was rich in eloquence. In particular (since humanists were normally better at Latin than they were at Greek) Cicero was considered to be the pattern of refined and copious discourse. In eloquence humanists found far more than an exclusively aesthetic quality. As an effective means of moving leaders or fellow citizens toward one political course or another, eloquence was akin to pure power. Humanists cultivated rhetoric, consequently, as the medium through which all other virtues could be communicated and fulfilled.

Humanism, then, may be accurately defined as that Renaissance movement which had as its central focus the ideal of *humanitas.* The narrower definition of the Italian term *umanisti* notwithstanding, all the Renaissance writers who cultivated *humanitas,* and all their direct "descendants," may be correctly termed humanists.

Other uses. It is small wonder that a term as broadly allusive as *humanism* should be subject to a wide variety of applications. Of these (excepting the historical movement described above) there are three basic types: humanism as classicism, humanism as referring to the modern concept of the humanities, and humanism as human-centredness.

Accepting the notion that Renaissance humanism was

simply a return to the classics, some historians and philologists have reasoned that classical revivals occurring anywhere in history should be called humanistic. St. Augustine, Alcuin, and the scholars of 12th-century Chartres have thus been referred to as humanists. In this sense the term can also be used self-consciously, as in the New Humanism movement in literary criticism led by Irving Babbitt and Paul Elmer More in the early 20th century.

The word *humanities,* which like the word *umanisti* derived from the Latin *studia humanitatis,* is often used to designate the nonscientific scholarly disciplines: language, literature, rhetoric, philosophy, art history, and so forth. Thus it is customary to refer to scholars in these fields as humanists and to their activities as humanistic.

Humanism and related terms are frequently applied to modern doctrines and techniques that are based on the centrality of human experience. In the 20th century the pragmatic humanism of Ferdinand C.S. Schiller, the Christian humanism of Jacques Maritain, and the movement known as secular humanism, though differing from each other significantly in content, all show this anthropocentric emphasis.

Not only is such a large assortment of definitions confusing, but the definitions themselves are often redundant or impertinent. There is no reason to call all classical revivals humanistic when the word classical suffices. To say that professors in the many disciplines known as the humanities are humanists is to compound vagueness with vagueness, for these disciplines have long since ceased to have or even aspire to a common rationale. The definition of humanism as anthropocentricity or human-centredness has a firmer claim to correctness. For obvious reasons, however, it is confusing to apply this word to classical literature.

BASIC PRINCIPLES AND ATTITUDES

Underlying the early expressions of humanism were principles and attitudes that gave the movement a unique character and would shape its future development.

Classicism. Early humanists returned to the classics less with nostalgia or awe than with a sense of deep familiarity, an impression of having been brought newly into contact with expressions of an intrinsic and permanent human reality. Petrarch, the acknowledged founder of the humanistic movement, dramatized his feeling of intimacy with the classics by writing "letters" to Cicero and Livy. Coluccio Salutati remarked with pleasure that possession of a copy of Cicero's letters would make it possible for him to talk with Cicero. Niccolò Machiavelli would later immortalize this experience in a letter that described his own reading habits in ritualistic terms:

'he
ursuit of
:ading

> Evenings I return home and enter my study; and at its entrance I take off my everyday clothes, full of mud and dust, and don royal and courtly garments; decorously reattired, I enter into the ancient sessions of ancient men. Received amicably by them, I partake of such food as is mine only and for which I was born. There, without shame, I speak with them and ask them about the reason for their actions; and they in their humanity respond to me.

Machiavelli's term *umanità* ("humanity") means more than kindness; it is a direct translation of the Latin *humanitas.* Machiavelli implies that he shared with the ancients a sovereign wisdom of human affairs. He also describes that theory of reading as an active and even aggressive pursuit that was common among humanists. Possessing a text and understanding its words were not enough; analytic ability and a questioning attitude were necessary before a reader could truly enter the councils of the great. These councils, moreover, were not merely serious and ennobling; they held secrets available only to the astute, secrets the knowledge of which could transform life from a chaotic miscellany into a crucially heroic experience. Classical thought offered insight into the heart of things. In addition, the classics suggested methods by which, once known, human reality could be transformed from an accident of history into an artifact of will. Antiquity was rich in examples, actual or poetic, of epic action, victorious eloquence, and applied understanding. Carefully studied and well employed, classical rhetoric could implement enlightened policy, while classical poetics could carry enlightenment into the very souls of men. In a manner that might seem paradoxical to more modern minds, humanists associated classicism with the future.

Realism. Early humanists shared in large part a realism that rejected traditional assumptions and aimed instead at the objective analysis of perceived experience. To humanism is owed the rise of modern social science, which emerged not as an academic discipline but rather as a practical instrument of social self-inquiry. Humanists avidly read history, taught it to their young, and, perhaps most importantly, wrote it themselves. They were confident that proper historical method, by extending across time their grasp of human reality, would enhance their active role in the present. For Machiavelli, who avowed to treat of men as they were and not as they ought to be, history would become the basis of a new political science. Similarly, direct experience took precedence over traditional wisdom. Leon Battista Alberti's dictum that an essential form of wisdom could be found only "at the public marketplace, in the theatre, and in people's homes" would be echoed by Francesco Guicciardini:

History and experience

> I, for my part, know no greater pleasure than listening to an old man of uncommon prudence speaking of public and political matters that he has not learnt from books of philosophers but from experience and action; for the latter are the only genuine methods of learning anything.

Renaissance realism also involved the unblinking examination of human uncertainty, folly, and immorality. Petrarch's honest investigation of his own doubts and mixed motives is born of the same impulse that led Giovanni Boccaccio in the *Decameron* to conduct an encyclopaedic survey of human vices and disorders. Similarly critical treatments of society from a humanistic perspective would be produced later by Erasmus, More, Castiglione, Rabelais, and Montaigne. But it was typical of humanism that this moral criticism did not, conversely, postulate an ideal of absolute purity. Humanists asserted the dignity of normal earthly activities and even endorsed the pursuit of fame and the acquisition of wealth. The emphasis on a mature and healthy balance between mind and body, first implicit in Boccaccio, is evident in the work of Giannozzo Manetti, Francesco Filelfo, and Paracelsus; it is embodied eloquently in Montaigne's final essay, "Of Experience." Humanistic tradition, rather than revolutionary inspiration, would lead Francis Bacon to assert in the early 17th century that the passions should become objects of systematic investigation. The realism of the humanists was, finally, brought to bear on the Roman Catholic Church, which they called into question not as a theological structure but as a political institution. Here as elsewhere, however, the intention was neither radical nor destructive. Humanism did not aim to remake humanity but rather to reform social order through an understanding of what was basically and inalienably human.

Critical scrutiny and concern with detail. Humanistic realism bespoke a comprehensively critical attitude. Indeed, the productions of early humanism constituted a manifesto of independence, at least in the secular world, from all preconceptions and all inherited programs. The same critical self-reliance shown by Coluccio Salutati in his textual emendations and Boccaccio in his interpretations of myth was evident in almost the whole range of humanistic endeavour. It was cognate with a new specificity, a profound concern with the precise details of perceived phenomena, that took hold across the arts and the literary and historical disciplines and would have profound effects on the rise of modern science. The increasing prominence of mathematics as an artistic principle and academic discipline was a testament to this development.

The emergence of the individual and the idea of the dignity of man. These attitudes took shape in concord with a sense of personal autonomy that first was evident in Petrarch and later came to characterize humanism as a whole. An intelligence capable of critical scrutiny and self-inquiry was by definition a free intelligence; the intellectual virtue that could analyze experience was an integral part of that more extensive virtue that could, according to many humanists, go far in conquering fortune. The

emergence of Renaissance individualism was not without its darker aspects. Petrarch and Alberti were alert to the sense of estrangement that accompanies intellectual and moral autonomy, while Machiavelli would depict, in *The Prince,* a grim world in which the individual must exploit the weakness of the crowd or fall victim to its indignities. But happy or sad, the experience of the individual had taken on a heroic tone. Parallel with individualism arose, as a favourite humanistic theme, the idea of the dignity of man. Backed by medieval sources but more sweeping and insistent in their approach, spokesmen such as Petrarch, Manetti, Valla, and Ficino asserted man's earthly preeminence and unique potentialities. In his noted *De hominis dignitate oratio* ("Oration on the Dignity of Man"), Giovanni Pico della Mirandola conveyed this notion with unprecedented vigour. Humanity, Pico asserted, had been assigned no fixed character or limit by God but instead was free to seek its own level and create its own future. No dignity, not even divinity itself, was forbidden to human aspiration. Pico's radical affirmation of human capacity shows the influence of Ficino's recent translations of the Hermetic writings. Together with the even bolder 16th-century formulations of this position by Paracelsus and Giordano Bruno, the *Oratio* betrays a rejection of the early humanists' emphasis on balance and moderation; it suggests the straining toward absolutes that would characterize major elements of later humanism.

Active virtue. The emphasis on virtuous action as the goal of learning was a founding principle of humanism and (though sometimes sharply challenged) continued to exert a strong influence throughout the course of the movement. Salutati, the learned chancellor of Florence whose words could batter cities, represented in word and deed the humanistic ideal of an armed wisdom: that combination of philosophical understanding and powerful rhetoric which alone could effect virtuous policy and reconcile the rival claims of action and contemplation. In *De ingenuis moribus et liberalibus studiis* ("On the Manners of a Gentleman and Liberal Studies"), a treatise that influenced Guarino Veronese and Vittorino da Feltre, Pietro Paolo Vergerio maintained that just and beneficent action was the purpose of humanistic education; his words were echoed by Alberti in *Della famiglia* ("On the Family"):

Armed wisdom

> As I have said, happiness cannot be gained without good works and just and righteous deeds. . . . The best works are those that benefit many people. Those are most virtuous, perhaps, that cannot be pursued without strength and nobility. We must give ourselves to manly effort, then, and follow the noblest pursuits.

Matteo Palmieri wrote that

> the true merit of virtue lies in effective action, and effective action is impossible without the faculties that are necessary for it. He who has nothing to give cannot be generous. And he who loves solitude can be neither just, nor strong, nor experienced in those things that are of importance in government and in the affairs of the majority.

Palmieri's philosophical poem, *La città di vita* ("The City of Life"), developed the idea that the world was divinely ordained to test human virtue in action. Later humanism would broaden and diversify the theme of active virtue. Machiavelli saw action not only as the goal of virtue but also (via historical understanding of great deeds of the past) as the basis for wisdom. Baldassare Castiglione, in his highly influential *Libro del cortegiano* (*Book of the Courtier*), developed in his ideal courtier a psychological model for active virtue, stressing moral awareness as a key element in just action. François Rabelais used the idea of active virtue as the basis for anticlerical satire. In his profusely humanistic *Gargantua,* he has the active hero Friar John save a monastery from enemy attack, while the monks sit uselessly in the church choir, chanting meaningless Latin syllables. John later asserts that, had he been present, he would have used his manly strength to save Jesus from crucifixion, and he castigates the Apostles for betraying Christ "after a good meal." Endorsements of active virtue, as will be shown, would also characterize the work of English humanists from Sir Thomas Elyot to John Milton. They typify the sense of social responsibility, the instinctive association of learning with politics and morality, that stood at the heart of the movement. As Salutati put it, "One must stand in the line of battle, engage in close combat, struggle for justice, for truth, for honour."

Rabelais's *Gargantua*

EARLY HISTORY

The influence of Petrarch (Francesco Petrarca, 1304–74) was profound and many-sided. As the most prominent man of letters of the 14th century, he promoted the recovery and transcription of classical texts, providing the impetus for the important classical researches of Boccaccio and Salutati. He threw himself into controversies in which he defined a new humanism in contradistinction to what he considered to be the barbaric influence of medieval tradition. He carried on an energetic correspondence that established him as a cultural focal point and would provide, if all his other works were lost, an accurate index of his views and their development. As a theologian (he was an ordained priest) he advanced the view, held by many humanists to follow, that classical learning and Christian spirituality were not only compatible but also mutually fulfilling. As a political apologist, he gave hearty support to Cola di Rienzo's brief revival of the Roman Republic (1347). As a poet, he was the first Renaissance writer to produce a Latin epic (*Africa*), but he was even more important for his compositions in the vernacular. His *Canzoniere* provided the model on which the Renaissance lyric was to take shape and the standard by which future productions would be judged. His work established secular poetry as a serious and noble pursuit. His eloquent and forceful presence made him a personal symbol of his own ideas. Crowned with laurel, favoured by rulers, legates, and scholars, he became the human focus for the new interest in classical revival and literary artistry.

Petrarch's greatest influence

It was, however, as a philosophical spokesman that Petrarch exerted his greatest influence on the history of humanism. In his prose works and letters he established many of the positions that would be central to the movement and broached many of the issues that would be its favourite subjects for debate. His idea of the poet as a philosophical teacher and thus as a champion of culture would inspire humanists from Boccaccio to Sidney. His endorsement of the study of rhetoric and his underlying notion of language as an informing principle of the individual and society would become crucial subjects of humanistic discussion and debate. His view of classical culture, not as an undifferentiated element of the past but as an authentic alternative to his own medieval society, was of equal historical importance. Petrarch broke with the past and helped to reestablish the Socratic tradition in Europe by specifying self-knowledge as a primary goal of philosophy. This attitude and his unfailing insistence on moral autonomy were early and important signs of the individualism that would become a Renaissance hallmark. He emphasized human virtue as opposed to fortune, thus setting the stage for numerous famous treatments of this theme. He struggled repeatedly with the dilemma of action versus contemplation, establishing it as a favourite topic for humanistic debate. Petrarch did not invent these subjects, nor does he usually treat them with overwhelming power. His preeminence lies in the fact that he was the first writer since antiquity to assert that they and other human matters were valid issues for philosophical inquiry in and of themselves, and in the energy and eloquence with which he made his work their forum.

Petrarch's influence was immediately apparent in the work of two major Florentine humanists, Giovanni Boccaccio and Coluccio Salutati. A close friend and devoted supporter of Petrarch, Boccaccio (1313–75) not only enlarged upon his preceptor's ideas but also made important humanistic contributions of his own. His *Teseide* was the first classical epic to have been written in the vernacular and influenced the more famous Italian epics of Ariosto and Tasso. His *De genealogia deorum gentilium* ("On the Genealogy of the Gods of the Gentiles"), a scholarly interpretive compendium of classical myth, was the first in a long line of Renaissance mythographies; it includes a celebrated defense of poetry as a medium of hidden truth, a stimulant to virtue, and a source of mental health. His most memorable contribution to humanism, however, was

ɔccaccio's *ecameron*

probably the famous *Decameron.* Ostensibly this work is no more than a collection of 100 tales about love. But subjected to the interpretive scrutiny that Boccaccio himself recommends in *De genealogia deorum gentilium,* the *Decameron* takes on a far more serious tone. The opening phrase "*Umana cosa è*" ("It is a human thing") is deeply thematic, reminding us that the author structured his work on Dante's spiritual epic, *La divina commedia.* A close reading of the *Decameron* suggests that in it Boccaccio is trying to establish for the human realm the same sort of comprehensive understanding that Dante established for the life of the spirit. Through moral fable and direct address to the reader, he undertakes a reinterpretation of human experience based not on traditional doctrine but rather on perceived reality. Appealing repeatedly to reason and nature, and constantly implying the superiority of awareness to innocence (which he equates with ignorance), he calls for a moral order built fairly and solidly on the potentialities of human nature. His 10 storytellers, who leave the plague-ravaged and chaotic city of Florence and reestablish themselves at a delightfully landscaped villa, suggest the remaking of culture through disentanglement with the past, unprejudiced analysis, and enlightened imagination. Rightly considered to be the wellspring of Western realism, the *Decameron* is also a monument to humanism. Though it makes little mention of classical thought, Boccaccio's great work rings with a tone that was even more basic to the humanistic movement: an emphasis on the human capacity for self-knowledge and willed renewal.

'oluccio alutati

Other humanistic elements implicit in Petrarch's thought were developed in the life and work of Coluccio Salutati (1331–1406). Like Petrarch, Salutati collected manuscripts, wrote on morality and politics, and carried on a voluminous correspondence. He was an aggressive and scientific philologist, instrumental in establishing principles of textual criticism that would become key elements of the humanistic method. He was a forceful apologist for the active life, and his theories bore fruit in his own career as chancellor of the Florentine republic. His use of classical eloquence in the service of his state was an early documentation of the humanistic faith in the political power of rhetoric; it led a bitter enemy, Gian Galeazzo Visconti of Milan, to say that a thousand Florentine horsemen had hurt him less than the letters of Coluccio. Salutati was succeeded in the Florentine chancellorship by two scholar-statesmen who reflected his influence, first Leonardo Bruni (1369–1444) and then Gian Francesco Poggio Braccioloni (1380–1459). Bruni was a pioneer in the advocacy of humanistic education, holding that the *studia humanitatis* shape the perfected man and that the goal of this perfected virtue is political action. His theory of education stressed the importance of practical experience (implicit in the work of Boccaccio) and put heavy emphasis on historical studies. His history of Florence is considered to be the first work of modern historiography; and, under the influence of Emmanuel Chrysoloras (1368–1415), a Byzantine teacher who had lectured at Florence and Pavia, he produced Latin translations of Plato and Aristotle that broke with medieval tradition by reproducing the sense of the Greek prose rather than following it word by word. Poggio, the foremost recoverer of classical texts, was also a moralist, a historian, a brilliant correspondent, and an early scholar of architectural antiquities. His long career, which included service to both church and state and friendships with Salutati, Bruni, Niccolò Niccoli, Guarino, Nicholas of Cusa, Donatello, and Cosimo de' Medici, exemplifies the scope and vitality of Italian humanism. Together these Florentine chancellors, whose active lives spanned almost a century, strengthened and consolidated the humanistic program. Moreover, their leadership strongly influenced the cultural developments that would make 15th-century Florence the most active intellectual and artistic centre in Europe.

As one proceeds with the history of humanism, the following major points about its development in the 14th century ought to be kept in mind. Humanism received its crucial imprint from the work of a single man and thence developed among men who maintained close touch with each other and acknowledged a shared mission. Humanism was not originally an academic movement but rather a program defined and promoted by statesmen and men of letters. Its proclaimed goal was widespread cultural renewal; therefore, it chose its subjects for consideration from the phenomena of human life as lived and adopted the Ciceronian model of philosopher as citizen in preference to the contemplative ideal. The heavy emphasis on civic action is connected with the fact that humanism developed in a republic rather than a monarchy.

The humanistic schools

By the turn of the 15th century, all of the key elements that came to define humanism were in place except for two: its detailed educational system and what might be called its Greek dimension. The founders of the first humanistic schools were Vittorino da Feltre (1373–1446) and Guarino Veronese (Guarino da Verona, 1374–1460). Vittorino and Guarino were fellow students at the University of Padua at the turn of the century; they are said later to have tutored each other (Guarino as an expert in Greek, Vittorino in Latin) after Guarino had opened the first humanistic school (Venice, *c.* 1414). Vittorino taught in both Padua (where he was briefly professor of rhetoric) and Venice during the early 1420s. In 1423 he accepted the invitation of Gianfrancesco Gonzaga, marquis of Mantua, to become tutor to the ruling family. At this post Vittorino spent the remaining 22 years of his life. His school, held in a delightful palace that he renamed "La Giocosa," had as its students not only the Gonzaga children (among them the future marquis, Ludovico) but also an increasing number of others, including sons of Poggio, Guarino, and Filelfo. The eminent humanist Lorenzo Valla studied there, as did Federico da Montefeltro, who later promoted humanistic institutions as duke of Urbino. Vittorino's school in Mantua was the first to focus the full power of the humanistic program, together with its implications in other arts and sciences, upon the education of the young. Latin literature, Latin composition, and Greek literature were required subjects of study. Heavy emphasis was placed on Roman history as an educational treasury of great men and memorable deeds. Rhetoric (as taught by Quintilian) was a central topic, not as an end in itself but as an effective means of channeling moral virtue into political action. Vittorino summed up the essentially political thrust of humanistic education as follows:

> Not everyone is called to be a physician, a lawyer, a philosopher, to live in the public eye, nor has everyone outstanding gifts of natural capacity, but all of us are created for the life of social duty, all are responsible for the personal influence that goes forth from us.

Other studies at Mantua included music, drawing, astronomy, and mathematics. The meadows around La Giocosa were turned into playing fields. Vittorino's educational policy spoke at once to mind and body, to aesthetic enjoyment and moral virtue. His work embodied a more comprehensive appeal to human perfectibility than had been attempted since antiquity. Humanists were not unaware of the originality and ambitiousness of this project. With reference to a similar program of his own, Guarino's son Battista remarked that "no branch of knowledge embraces so wide a range of subjects as that learning that I have now attempted to describe."

Byzantine influence

Guarino had learned his Greek in Constantinople under the influence of Chrysoloras, whose dynamic presence had done much to foster Greek studies in Italy. During the course of the 15th century, which saw the famous council of Eastern and Western churches (Ferrara–Florence, 1438–45) and later the fall of Constantinople to the Turks (1453), Italy received as welcome immigrants a number of other eminent Byzantine scholars. George Gemistus Plethon (1355–1450) was a major force in Cosimo de' Medici's foundation of the Platonic Academy of Florence. George of Trebizond (Georgius Trapezuntius, 1395–1484), a student of Vittorino, was a formidable bilingual stylist who wrote important handbooks on logic and rhetoric. Theodore Gaza (*c.* 1400–75) and Johannes Argyropoulos (1410–90) contributed major translations of Aristotle. John (originally Basil) Bessarion (1403–72), who became a cardinal in 1439, explored theology from a Platonic perspective and sought to resolve apparent conflicts between Platonic and Aristotelian philosophy; his large collection

of Greek manuscripts, donated to the Venetian senate, became the core of the notable library of St. Mark. This infusion of Byzantine scholarship had a profound effect on Italian humanism. By making Greek texts and commentaries available to Western students, and by acquainting them with Byzantine methods of criticism and interpretation, the teachers from Constantinople enabled Italian humanists to explore the bases of classical thought and to appreciate its greatest monuments, either in the original or in accurate new Latin translations.

THE 15TH CENTURY

As Italian humanism grew in influence during the 15th century, it developed ramifications that connected it with every major field of intellectual and artistic activity. Moreover, the advent of printing at mid-century and the contemporaneous upsurge of publication in the vernacular brought new sectors of society under humanistic influence. These and other cultural impetuses hastened the export of humanistic ideas to the Low Countries, France, England, and Spain, where significant humanistic programs would be in place by the early 16th century. Even as these things were happening, however, other changes were deeply and permanently affecting the character of the movement. The concerns of many major humanists were narrowed by inevitable historical processes of specialization, to the extent that, in a large number of cases, humanism lost its comprehensive thrust and became a predominantly academic or literary pursuit. The political élan of humanism was weakened by the decline of republican institutions in Florence. Ambiguities and paradoxes implicit in the original program developed into open conflicts, dividing the movement into camps and depleting much of its original integrity. But before considering these developments, one might do well to appreciate three 15th-century examples of humanism at its height: the career of Leon Battista Alberti and the humanistic courts at Florence and Urbino.

Leon Battista Alberti. The achievement of Leon Battista Alberti (1404–72) testifies to the formative power and exhaustive scope of earlier Italian humanism. He owed his boyhood education to Gasparino da Barzizza (1359–1431), the noted teacher who, with Vergerio, was influential in the development of humanism at Padua. Alberti attended the University of Bologna from 1421 until 1428, by which time he was expert in law and mathematics and so adept at humanistic literary skills that his comedy *Philodoxeos* was accepted as the newly discovered work of an ancient author. In 1428 he became secretary to Cardinal Albergati, bishop of Bologna, and in 1432 he accepted a similar position in the papal chancery at Rome. His service to the church soon brought him incomes that permanently secured his livelihood, and he spent the remainder of his life at a variety of literary, philosophical, and artistic pursuits so dazzling as to challenge belief. He was a poet, essayist, and biographer. His moral and philosophical works, especially *Della famiglia, De iciarchia* ("On the Man of Excellence and Ruler of His Family"), and *Momus,* are humanistic statements that nonetheless bear the mark of a unique individual. He wrote a rhetorical handbook and a grammatical treatise, the *Regule lingue Florentine,* which bespeaks his strong influence on the rise of literary expression in the vernacular. He contributed an important text on cartography and was instrumental in the development of ciphers. A prominent architect (*e.g.,* the Tempio Malatestiano in Rimini and the facade of Sta. Maria Novella in Florence), he was also an eminent student of all artistic ideas and practices. His three studies—*De pictura* (*On Painting*), *De statua* (*On Sculpture*), and *De re aedificatoria* (*Ten Books on Architecture*)—were landmarks in art theory, powerful in developing the theory of perspective and the idea of "human" space. His theoretical and practical reliance on mathematics (which he considered to be the basic, unifying element of all science) is rightly seen as an important step in the early development of modern method.

Alberti's writings

Behind these achievements was a man of startling physical prowess and inexhaustible sanguinity. He said outright that an individual could encompass whatever project he truly willed, and his own life bore witness to this radical thesis. In the 19th century Jacob Burckhardt would write of him as a "universal man" of the Renaissance, while his own contemporary Politian described him with wonderment: "It is better to be silent about him than not to say enough." Alberti's theory and practice bore an undeniably humanistic stamp. His passion for mathematics was in all likelihood an outgrowth of the educational program at Padua (Vittorino, himself an avid mathematician, was also a student of Barzizza). His omnivorous pursuit of knowledge recalls Barzizza's conviction that *humanitas* was the unifying principle of many arts. An advocate of classical erudition in art and architecture as well as in literary activity, he extended into his artistic studies the same sense of precision and specificity that earlier humanists had applied to philology. His sense of human dignity, evident in all his productions, was supported and indeed justified by a strenuous realism. His advocacy of the vernacular disturbed a number of more doctrinaire humanists, who favoured total Latinity. But this predisposition, rather than a divergence from humanistic principle, was a direct outgrowth of its evangelistic thrust. In short, Alberti uniquely fulfilled the humanistic aspiration for a learning that would comprehend all experience and a philosophical heroism that would renew society.

The Medici and Federico da Montefeltro. The 15th century saw the rise of the Platonic Academy of Florence and the great humanistic courts. Close ties between Poggio and the Medici helped make that ruling family of Florence the new custodians of the humanistic heritage. Cosimo de' Medici (Cosimo the Elder, 1389–1464), who had personally lured the great council of churches from Ferrara to Florence in 1439, became so enamoured of Greek learning that, at the suggestion of Gemistus Plethon, he decided to found a Platonic academy of his own. He amassed a great collection of books, which would form the nucleus of the Laurentian Library. He generously supported the work of scholars, in particular encouraging the brilliant Marsilio Ficino (1433–99) to undertake a complete Latin translation of Plato. Other notable members of the academy were Politian, Cristoforo Landino (1424–1504), and Ficino's own student, Giovanni Pico della Mirandola (1463–94). The Medici family was equally notable in its patronage of the arts, supporting projects by a list of masters that included Brunelleschi, Michelangelo, and Cellini. Cosimo's famous grandson Lorenzo (Lorenzo the Magnificent, 1449–92) was of a thoroughly humanistic disposition. Lorenzo's versatile and energetic nature lent itself equally to politics and philosophy, to martial arts and music. He wrote poetry and literary commentary and formed close ties with Ficino, Pico, and other leading scholars of the academy. He continued his grandfather's lavish patronage of art and learning and was said to have spent half of his city's revenues on the purchase of books alone. Active in many fields, he nonetheless acknowledged the preeminence of the life of the mind. When chided by a friend for sleeping late and not going out to work, Lorenzo replied, "What I have dreamed in one hour is worth more than what you have done in four."

The Medici patronage

The influence of humanism was evident in many 15th-century Italian courts, including Rome itself, which boasted, in Pius II (Enea Silvio Piccolomini, also known as Aeneas Sylvius Piccolomini, 1405–64), a humanist pope. It manifested itself strikingly at Urbino, where Federico da Montefeltro (1422–82) turned an isolated hill town into a treasury of Renaissance culture. Schooled by Vittorino in Mantua, Federico chose warfare as his calling. As a mercenary he gained a reputation for winning his battles and keeping his word, and the fortune he accumulated in fees and prizes became the medium for his city's renewal. He brought architects, artists, and scholars to Urbino and built a great palace whose unadorned exterior concealed magnificent chambers, a graceful courtyard, and a secret garden. Federico was enthusiastically devoted to the collection and preservation of books. His library, described by Vespasiano Bisticci as being even more complete than that of the Medici, contained an army of 30 to 40 scribes who were constantly at work. His own virtues were so notable and diverse as to mark him as a possible model for Rabelais's humanistic giant, Gargantua. Mighty at

Federico's library

arms, he was also conscientious in religious observances; supremely powerful, he was nonetheless a modest and courteous companion. Beneath the ivied tranquility of his secret garden stretched an indoor equestrian arena. He commissioned paintings by Piero della Francesca and was the object of humanistic dedications by Poggio, Landino, and Ficino. He kept two organists at court and maintained five men to read the classics aloud at meals. Federico's intellectual accomplishments were impressive. His skill at mathematics shows the influence of Vittorino. He was a good Latinist and as a student of classical history was able to hold his own in conversation with the erudite Pius II. At philosophy Federico was even more astute. Vespasiano wrote that

> he began to study logic with the keenest understanding, and he argued with the most nimble wit that was ever seen. After he had heard (Aristotle's) *Ethics* many times, comprehending it so thoroughly that his teachers found him hard to cope with in disputation, he studied the *Politics* assiduously. . . . Indeed, it may be said of him that he was the first of the Signori who took up philosophy and had knowledge of the same. He was ever careful to keep intellect and virtue to the front, and to learn some new thing every day.

Federico's balance and versatility made him, even more than Lorenzo, an example of the humanistic program in action. Baldassare Castiglione, perhaps the most thoughtful of the later Italian humanists, would speak of him as "the light of Italy; there is no lack of living witnesses to his prudence, humanity (*umanità*), justice, intrepid spirit, (and) military discipline." Castiglione described Federico's residence as seeming to be less a palace than "a city in the form of a palace"; one might say as well that this structure, with its elegant accommodation for every creative human activity, was an architectural image of the humanistic mind.

LATER ITALIAN HUMANISM

The achievement of Alberti, Federico, and the Medici up to Lorenzo may be seen as the effective culmination of Italian humanism, the ultimate realization of its motives and principles. At the same time as these goals were being achieved, however, the movement was beginning to suffer bifurcation and dilution. Even the enthusiastic Platonism of the Florentine academy was, in its idealism and emphasis on contemplation, a significant digression from the crucial humanistic doctrine of active virtue, and Pico della Mirandola himself was politely admonished by a friend to forsake the ivory tower and accept his civic responsibilities. The conflicting extremes to which sincere humanistic inquiry could drive scholars are nowhere more apparent than in the fact that the arch-idealist Pico and the arch-realist Machiavelli lived in the same town and at the same time. Castiglione, who had belonged to the court of Federico's son Guidobaldo, would be saddened by its decline and shocked when another of his patrons, the "model" Renaissance prince Charles V, ordered the sack of Rome. To a large extent, the cause of these and other vicissitudes lay in the nature of the movement itself, for that boundless diversity which nourished its strength was also a well of potential conflict. Humanists' undifferentiated acceptance of the classical heritage was also in effect an appropriation of the profound controversy implicit in that heritage. Rifts between Platonists, monarchists, and republicans; positivists and skeptics; idealists and cynics; and historians and poets came to be more and more characteristic of humanistic discourse. Some of these tensions had been clear from the start, Petrarch having been ambiguous in his sentiments regarding action versus contemplation, and Salutati having been not wholly clear about whether he preferred republics to monarchies. But the 15th century, bringing with it the irreconcilable heterogeneity of Greek thought, vastly multiplied and deepened these divisions. Of these schisms, the two that perhaps most deeply influenced the course of humanism were the so-called *res–verbum* ("thing–word") controversy and the split between Platonic idealism and historical realism.

Diversification and erosion

Things and words. Simply put, the *res–verbum* controversy was an extended argument between humanists who believed that language constituted the ultimate human reality and those who believed that language, though an important subject for study, was the medium for understanding an even more basic reality that lay beyond it. The origin of the controversy lay in the debate in the 5th–4th century BC between the Socratic school, which held that language was an important means of understanding deeper truths, and the Sophistic-rhetorical school, which held that "truth" was itself a fiction dependent on varying human beliefs and therefore that language had to be considered the ultimate arbiter. Petrarch, who had no direct contact with the works of Plato and little detailed knowledge of his ideas, drew on Cicero and St. Augustine in his development of a Christian-rhetorical position, holding that "it is more satisfying (*satius*) to will the good than to know the truth" and espousing rhetoric as the effective means of convincing people "to will the good."

Origin of the *res–verbum* controversy

This assertion would critically shape the character of humanism through the Renaissance and beyond. It was never effectively challenged by Renaissance Platonists because, for reasons discussed below, Renaissance Platonists, though strong in Platonic idealism, were weak in Platonic analytical method. The enthronement of language as both subject and object of humanistic inquiry is evident in the important work of Lorenzo Valla (1407–57) and Politian (Angelo Poliziano, 1454–94). Valla spoke of language as a "sacrament" and urged that it be studied scientifically and historically as the synthesis of all human thought. For Valla, the study of language was, in effect, the study of humanity. Similarly, Politian held that there were in fact two dialectics: one of ideas and one of words. Rejecting the dialectic of ideas as being too difficult and abstruse, he espoused the dialectic of words (*i.e.,* philology and rhetoric) as the proper human study. This project would bear fruit in the intensive linguistic-philosophical researches of Mario Nizolio (1498–1575). Though anticipated by Petrarch, the radical emphasis on the primacy of the word constituted a break with the teaching of other early humanists, such as Bruni and Vittorino, who had strongly maintained that the word was of value only through its relationship to perceived reality. Nor did the old viewpoint lack later adherents. In an epistolary debate with Ermolao Barbaro (1454–93), Pico asserted the preeminence of things over words and hence of philosophy over rhetoric: "But if the rightness of names depends on the nature of things, is it the rhetorician we ought to consult about this rightness, or is it the philosopher who alone contemplates and explores the nature of everything?" Appeals of this sort, however, were not to win the day. Philosophical humanism declined because, though rich in conviction, it had failed to establish a systematic relationship between philosophy and rhetoric, between words and things. By the 16th century, Italian humanism was primarily a literary pursuit, and philosophy was left to develop on its own. Despite significant challenges, the division between philosophical and literary studies would solidify in the development of Western culture.

Idealism and the Platonic Academy of Florence. The idealism so prominent in the Florentine academy is called Platonic because of its debt to Plato's theory of Ideas and to the epistemological doctrine established in his *Symposium* and *Republic.* It did not, however, constitute a complete appreciation or reassertion of Plato's thought. Conspicuously absent from the Florentine agenda was the analytic method (dialectic), which was Socrates' greatest contribution to philosophy. This major omission cannot be explained philologically, at least after Ficino's work had made the complete Platonic corpus available in clear Latin prose. The explanation lies rather in a specific cast of mind and in a dramatically successful forgery. The major Platonists of the mid-15th century, Plethon, Bessarion, and Nicholas of Cusa (Nicholaus Cusanus, 1401–64), had all concentrated their attention on the religious implications of Platonic thought; and, following them, Marsilio Ficino (1433–99) sought to reconcile Plato with Christ in a *pia philosophia* ("pious philosophy"). The transcendental goals of these philosophers left little room for the painstaking dialectical method that sifted through the details of perception and language, even though Plato himself had repeatedly alleged that transcendence itself was impos-

sible without this method. Along with Plato, moreover, Ficino had translated into Latin the works of the so-called Hermes Trismegistos. These books, which also emphasized transcendence at the expense of method, laid claim to divine authority and to an antiquity far greater than Plato's. They were, in fact, forgeries from a much later period, and are in many ways typical of the idealized and diluted versions of Plato that are called Neoplatonic. But the academy, and for that matter all the other Platonists of the 15th century, bought them wholesale. The result of these factors was a Platonism sans Platonic method, a philosophy that, straining for absolutes, had little interest in establishing its own basis in reality. Near the end of *The Book of the Courtier,* Castiglione puts a speech typical of Florentine Platonism in the mouth of his friend, the Platonist Pietro Bembo (1470–1547). As Bembo finishes his oration, a female companion tugs at the hem of his robe and says, "Take care, Master Pietro, that with such thoughts your soul does not forsake your body."

Machiavelli's realism. Niccolò Machiavelli (1469–1527), whose work derived from sources as authentically humanistic as those of Ficino, proceeded along a wholly opposite course. A throwback to the chancellor-humanists Salutati, Bruni, and Poggio, he served Florence in a similar capacity and with equal fidelity, using his erudition and eloquence in a civic cause. Like Vittorino and other early humanists, he believed in the centrality of historical studies, and he performed a signally humanistic function by creating, in *La Mandragola,* the first vernacular imitation of Roman comedy. His characteristic reminders of human weakness suggest the influence of Boccaccio; and like Boccaccio he used these reminders less as satire than as practical gauges of human nature. In one way at least, Machiavelli is more humanistic (*i.e.,* closer to the classics) than the other humanists, for while Vittorino and his school ransacked history for examples of virtue, Machiavelli (true to the spirit of Polybius, Livy, Plutarch, and Tacitus) embraced all of history, good, evil, and indifferent, as his school of reality. Like Salutati, though perhaps with greater self-awareness, Machiavelli was ambiguous as to the relative merits of republics and monarchies. In both public and private writings (especially the *Discorsi sopra la prima deca di Tito Livio* ["Discourses on the First Ten Books of Livy"]) he showed a marked preference for republican government, while in *The Prince* he developed, with apparent approval, a model of radical autocracy. For this reason, his goals have remained unclear.

Machiavelli's writings

His methods, on the other hand, were coherent throughout and remain a major contribution to social science and the history of ideas. Like earlier humanists, Machiavelli saw history as a source of power, but, unlike them (and here perhaps influenced by Sophistic and Averroistic thought), he saw neither history nor power itself within a moral context. Rather he sought to examine history and power in an amoral and hence (to him) wholly scientific manner. He examined human events in the same way that Alberti, Galileo, and the new science examined physical events: as discrete phenomena that had to be measured and described before they could be explained and evaluated. To this extent his work, though original in its specific design, was firmly based in the humanistic tradition. At the same time, however, Machiavelli's achievement significantly eroded humanism. By laying the foundations of modern social science, he created a discipline that, though true to humanistic methodology, had not the slightest regard for humanistic morality. In so doing, he brought to the surface a contradiction that had been implicit in humanism all along: the dichotomy between critical objectivity and moral evangelism.

The achievement of Castiglione. Though Italian humanism was being torn apart by the natural development of its own basic motives, it did not thereby lose its native attractions. The humanistic experience, in both its positive and negative effects, would be reenacted abroad. Baldassare Castiglione (1478–1529), whose *Book of the Courtier* affectionately summed up humanistic thought, was one of its most powerful ambassadors. Alert to the major contradictions of the program, yet intensely appreciative of its brilliance and energy, Castiglione wove its various strains together in a long dialogue that aimed at an equipoise between various humanistic extremes. Ostensibly a treatise on the model courtier, *The Book of the Courtier* is more seriously a philosophically organized pattern of conflicting viewpoints in which various positions—Platonist and Aristotelian, idealist and cynic, monarchist and republican, traditional and revolutionary—are given eloquent expression. Unlike most of his humanistic forebears, Castiglione is neither missionary nor polemical. His work is not an effort at systematic knowledge but rather an essay in higher discretion, a powerful reminder that every virtue (moral or intellectual) suggests a concomitant weakness and that extreme postures tend to generate their own opposites. The structure of the dialogue, in which Bembo's Platonic ecstasy is balanced by Bibbiena's assortment of earthy jests, is a testament to this intention. While Castiglione's professed subject matter would epidemically inspire European letters and manners of the 16th century, his more profound contribution would be echoed in the work of Montaigne and Shakespeare. His work suggests a redefined humanism, a virtue matured in irony and directed less toward knowledge than toward wisdom.

The Book of the Courtier

Tasso's Aristotelianism. In 16th-century Italy, humanistic methods and attitudes provided the medium for a kaleidoscopic variety of literary and philosophical productions. Of these, the work that perhaps most truly reflected the original spirit of humanism was the *Gerusalemme liberata* of Torquato Tasso (1544–95). New humanistic translations of Aristotle during the 15th century had inspired an Aristotelian Renaissance, and the attention of literary scholars focused particularly on the *Poetics.* In constructing his epic poem, Tasso was strongly influenced by Aristotle's views regarding the philosophical dimension of poetry; loosely paraphrasing Aristotle, he held (in his *Apologia*) that poetry, by incorporating both particulars and universals, was capable of seeking truth in its perfect wholeness. As a vehicle for philosophical truth, poetry consequently could provide moral education, specifically in such virtues (reinterpreted from a Christian perspective) as Aristotle had described in the *Nichomachean Ethics.* The Aristotelian Renaissance thus facilitated the revival of one of the chief articles in the original humanistic constitution: the belief in the poet's role as renewer of culture.

NORTHERN HUMANISM

Though humanism in northern Europe and England sprang largely from Italian sources, it did not emerge exclusively as an outgrowth of later Italian humanism. Non-Italian scholars and poets found inspiration in the full sweep of the Italian tradition, choosing their sources from Petrarch to Castiglione and beyond.

Desiderius Erasmus. Erasmus (*c.* 1466–1536) was the only other humanist whose international fame in his own time compared with Petrarch's. While lacking Petrarch's polemical zeal and spirit of self-inquiry, he shared the Italian's intense love of language, his dislike for the complexities and pretenses of medieval institutions both secular and religious, and his commanding personal presence. More specifically, however, his ideas and overall direction betray the influence of Lorenzo Valla, whose works he treasured. Like Valla, who had attacked biblical textual criticism with a vengeance and proved the so-called Donation of Constantine to be a forgery, Erasmus contributed importantly to Christian philology. Also like Valla, he philosophically espoused a kind of Christian hedonism, justifying earthly pleasure from a religious perspective. But he was most like Valla (and indeed the entire rhetorical "arm" of Italian humanism) in giving philology prominence over philosophy. He described himself as a poet and orator rather than an inquirer after truth. His one major philosophical effort, a Christian defense of free will, was thunderously answered by Luther. Though his writings are a well of good sense, they are seldom profound and are predominantly derivative. In Latin eloquence, on the other hand, he was preeminent, both as stylist and theorist. His graceful and abundant Ciceronian prose (whose principles he set down in *De copia verborum et rerum*) helped shape the character of European style. Perhaps his most original work is *Moriae encomium* (*The Praise of*

The Praise of Folly

Folly), an elegant combination of satire and poetic insight whose influence was soon apparent in the work of More (to whom it was dedicated) and Rabelais.

The French humanists. Erasmus' associates in France included the influential humanists Robert Gaguin (1433–1501), Jacques Lefèvre d'Étaples (*c.* 1455–1536), and Guillaume Budé (Guglielmus Budaeus, 1467–1540). Of these three, Budé was most central to the development of French humanism, not only in his historical and philological studies but also in his use of his national influence to establish the Collège de France and the library at Fontainebleau. The influence of Francis I (1494–1547) and his learned sister Margaret of Angoulême (1492–1549) was important in fostering the new learning. The diversity and energy of French humanism is apparent in the activities of the Estienne family of publishers; the poetry of Pierre de Ronsard (1524–85), Joachim du Bellay (*c.* 1522–60), and Guillaume du Bartas (1544–90); the political philosophy of Jean Bodin (1530–96); the philosophical methodology of Petrus Ramus (Pierre de la Ramée, 1515–72); and the dynamic relationship between humanistic scholarship and church reform (see below, *Humanism and Christianity*). Hampered by religious repression and compressed more severely in time, the French movement lacked the intellectual fecundity and the programmatic unity of its Italian counterpart. In François Rabelais and Michel de Montaigne, however, the development of humanistic methods and themes resulted in unique and memorable achievement.

François Rabelais (c. 1490–1533). Rabelais ranks with Boccaccio as a founding father of Western realism. As a satirist and stylist (in his hands French prose became a free, poetic form), he influenced writers as important as Jonathan Swift, Laurence Sterne, and James Joyce and may be seen as a major precursor of modernism. His five
Gargantua books concerning the deeds of the giant princes Gargantua and Pantagruel constitute a treasury of social criticism, an articulate statement of humanistic values, and a forceful, if often outrageous, manifesto of human rights. Rabelaisian satire took aim at every social institution and (especially in Book III) every intellectual discipline. Broadly learned and unflaggingly alert to jargon and sham, he repeatedly focused on dogmas that fetter creativity, institutional structures that reward hypocrisy, educational traditions that inspire laziness, and philosophical methodologies that obscure elemental reality. His heroes, Gargantua and his son and heir Pantagruel, are figures whose colossal size and appetites (Rabelais's etymology for *Pantagruel* is "all-thirsty") symbolize the nobility and omnivorous curiosity that typified the humanistic scheme. The multifarious educational program detailed in *Gargantua* is reminiscent of Vittorino, Alberti, and the Montefeltro court; and the utopian Abbey of Thélème, whose gate bears the motto "Do as you please," is a tribute to enlightened will and pleasure in the manner of Valla, Erasmus, and More. Characteristically overstated and never wholly free of irony, Rabelais's work is a far cry from the earnest moral and educational programs of the early humanists. Rather than rebuild society, he seeks to amuse, edify, and refine it. His qualified endorsement of human dignity is based on the healthy balance of mind and body, the sanctity of all true learning, and the authenticity of direct experience.

Michel de Montaigne (1533–92). Montaigne's famous *Essays* are not only a compendious restatement and reevaluation of humanistic motives but also a milestone in the humanistic project of self-inquiry that had been originally endorsed by Petrarch. Scholar, traveler, soldier, and statesman, Montaigne was, like Machiavelli, alert to both theory and practice; but while Machiavelli saw practice as forming the basis for sound theory, Montaigne perceived in human events a multiplicity so overwhelming as to deny theoretical analysis. Montaigne's use of typical humanistic modalities—interpretation of the classics, appeals to direct experience, exclusive emphasis on the human realm, and universal curiosity—led him, in other words, to the refutation of a typical humanistic premise: that knowledge of the intellectual arts could teach one a sovereign art of life. In an effort to make his inquiry more inclusive and unsparing, Montaigne made himself the subject of his book, demonstrating through hundreds of personal anecdotes and admissions the ineluctable diversity of a single human spirit. His essays, which seem to move freely from one subject or viewpoint to another, are often in fact carefully organized dialectical structures that draw the reader, through thesis and antithesis, stated subject and relevant association, toward a multidimensional understanding of morality and history. The final essay, grandly titled "Of Experience," counsels a mature acceptance of life in all its contradictions. Human dignity, he implies, is indeed possible, but it lies less in heroic achievement than in painfully won self-knowledge. In this sense Montaigne's attitude toward the humanistic tradition is generally similar to that suggested in the work of Castiglione and Rabelais. While effectively taking issue with a number of the more extreme humanistic contentions, he retained and indeed justified the basic attitudes that gave the movement its form.

"Of Experience"

The English humanists. English humanism flourished in two stages: the first a basically academic movement that had its roots in the 15th century and culminated in the work of Sir Thomas More, Sir Thomas Elyot, and Roger Ascham, the second a poetic revolution led by Sir Philip Sidney and William Shakespeare.

Though continental humanists had held court positions since the days of Humphrey of Gloucester, English humanism as a distinct phenomenon did not emerge until late in the 15th century. At Oxford William Grocyn (*c.* 1446–1519) and his student Thomas Linacre (*c.* 1460–1524) gave impetus to a tradition of classical studies that would permanently influence English culture. Grocyn and Linacre attended Politian's lectures at the Platonic Academy of Florence. Returning to Oxford, they became central figures in a group that included such younger scholars as John Colet (1466/67–1519) and William Lily (1468?–1522). The humanistic contributions of the Oxford group were philological and institutional rather than philosophical or literary. Grocyn lectured on Greek and theology; Linacre produced several works on Latin grammar and translated Galen into Latin. To Linacre is owed the foundation of the Royal College of Physicians; to Colet, the foundation of St. Paul's School, London. Colet collaborated with Lily (the first headmaster of St. Paul's) and Erasmus in writing the school's constitution, and together the three scholars produced a Latin grammar (known alternately as "Lily's Grammar" and the "Eton Grammar") that would be central to English education for decades to come.

In Sir Thomas More (1478–1535), Sir Thomas Elyot (*c.* 1490–1546), and Roger Ascham (1515–68), English humanism bore fruit in major literary achievement. Educated at Oxford (where he read Greek with Linacre), More was also influenced by Erasmus, who wrote *The Praise of Folly* (Latin *Moriae encomium*) at More's house and named the book punningly after his English friend. More's *Utopia* famous *Utopia,* a kind of companion piece to *The Praise of Folly,* is similarly satirical of traditional institutions (Book I) but offers, as an imaginary alternative, a model society based on reason and nature (Book II). Reminiscent of Erasmus and Valla, More's Utopians eschew the rigorous cultivation of virtue and enjoy moderate pleasures, believing that "Nature herself prescribes a life of joy (that is, pleasure)" and seeing no contradiction between earthly enjoyment and religious piety. Significantly indebted both to classical thought and European humanism, the *Utopia* is also humanistic in its implied thesis that politics begins and ends with humanity: that politics is based exclusively on human nature and aimed exclusively at human happiness. Sir Thomas Elyot chose a narrower subject but developed it in more detail. His great work, *The Book Named The Governor,* is a lengthy treatise on the virtues to be cultivated by statesmen. Born of the same tradition that produced *The Prince* and *The Courtier, The Governor* is typical of English humanism in its emphasis on the accommodation of both classical and Christian virtues within a single moral view. Elyot's other contributions to English humanism include philosophical dialogues, moral essays, translations of ancient and contemporary writers (including Isocrates and Pico), an important Latin-English dictionary, and a highly popular health manual. He served

his country as ambassador to the court of Charles V. Finally, the humanistic educational program set up at the turn of the century was vigorously supported by Sir John Cheke (1514–57) and codified by his student Roger Ascham. Ascham's famous pedagogical manual, *The Schoolmaster,* offers not only a complete program of humanistic education but also an evocation of the ideals toward which that education was directed.

Elizabeth I

Ascham had been tutor to the young princess Elizabeth, whose personal education was a model of humanistic pedagogy and whose writings and patronage bespoke great love of learning. Elizabeth I's reign (1558–1603) saw the last concerted expression of humanistic ideas. Elizabethan humanism, which added a unique element to the history of the movement, was the product not of pedagogues and philologists but of poets and playwrights.

Sidney and Spenser. Sir Philip Sidney (1554–86) was, like Alberti and Federico da Montefeltro, a living pattern of the humanistic ideal. Splendidly educated in the Latin classics at Shrewsbury and Oxford, Sidney continued his studies under the direction of the prominent French scholar Hubert Languet and was tutored in science by the learned John Dee. His brief career as writer, statesman, and soldier was of such acknowledged brilliance as to make him, after his tragic death in battle, the subject of an Elizabethan heroic cult. Sidney's major works, *Astrophel and Stella,* the *Defence of Poesie,* and the two versions of the *Arcadia,* are medleys of humanistic themes. In the sonnet sequence *Astrophel and Stella,* he surpassed earlier imitators of Petrarch by emulating not only the Italian humanist's subject and style but also his philosophical bent and habit of self-scrutiny. The *Defence of Poesie,* composed (like Erasmus' *Praise of Folly*) in the form of a classical oration, reasserts the theory of poetry as moral doctrine that had been articulated by Petrarch and Boccaccio and revived by the Italian Aristotelians of the 16th century. The later or "new" *Arcadia* is an epic novel whose theoretical concerns include the dualities of contemplation and action, reason and passion, and theory and practice. In this ambitious and unfinished work, Sidney attempts a characteristically humanistic synthesis of classical philosophy, Christian doctrine, psychological realism, and practical politics. Seen as a whole, moreover, Sidney's life and work form a significant contribution to a debate that had been smoldering since the decline of political liberty in Florence in the 15th century. How, it was asked, could humanism be politically active or "civic" in a Europe that was almost exclusively monarchic in structure? Many humanists had counseled retirement from active life, while Castiglione had seen his learned courtier rather as an advisor than as a leader. Sidney and his friend Edmund Spenser (1552/53–1599) sought to resolve this dilemma by creating a form of chivalric humanism. The image (taken on personally by Sidney and elaborated upon by Spenser in *The Faerie Queene*) of the hero as questing knight suggests that the humanist, even if not empowered politically, can achieve a valid form of activism by refining, upholding, and representing the values of a just and noble court. Spenser's poetic development of this humanistic program was even more specific than Sidney's. In his famous letter to Raleigh, he asserts that his purpose in *The Faerie Queene* is "to fashion a gentleman or noble person in virtuous and gentle discipline" and describes a project (never to be completed) of presenting his idea of the Aristotelian virtues in twelve poetic books. As with Sidney, however, this moral didacticism is neither self-righteous nor pedantic. The prescriptive content of *The Faerie Queene* is qualified by a strong emphasis on moral autonomy and a mature sense of the ambiguity of experience.

Chapman, Jonson, and Shakespeare. The poetry and drama of Shakespeare's time were a concourse of themes, ancient and modern, continental and English. Prominent among these motives were the characteristic topics of humanism. George Chapman (1559?–1634), the translator of Homer, was a forthright exponent of the theory of poetry as moral wisdom, holding that it surpassed all other intellectual pursuits. Ben Jonson (1572–1637) described his own humanistic mission when he wrote that a good poet was able "to inform young men to all good disciplines, inflame grown men to all great virtues, keep old men in their best and supreme state, or, as they decline to childhood, recover them to their first strength" and that the poet was "the interpreter and arbiter of nature, a teacher of things divine no less than human, a master in manners." Jonson, who sought this moral goal both in his tragedies and in his comedies, paid tribute to the humanistic tradition in *Catiline,* a tragedy in which Cicero's civic eloquence is portrayed in heroic terms.

Shakespeare

Less overtly humanistic, though in fact more profoundly so, was William Shakespeare (1564–1616). Thoroughly versed (probably at his grammar school) in classical poetic and rhetorical practice, Shakespeare early in his career produced strikingly effective imitations of Ovid and Plautus (*Venus and Adonis* and *The Comedy of Errors,* respectively) and drew on Ovid and Livy for his poem *The Rape of Lucrece.* In *Julius Caesar, Antony and Cleopatra,* and *Coriolanus* he developed Plutarchan biography into drama that, though Elizabethan in structure, is sharply classical in tone. Shakespeare clearly did not accept all the precepts of English humanism at face value. He grappled repeatedly with the problem of reconciling Christian doctrine with effective political action, and for a while (*e.g.,* in *Henry V*) seemed inclined toward the Machiavellian alternative. In *Troilus and Cressida,* moreover, he broadly satirized Chapman's Homeric revival and, more generally, the humanistic habit of idolizing classical heroism. Finally, he eschewed the moralism, rationalism, and self-conscious erudition of the humanists and was lacking as well in their fraternalism and their theoretical bent. Yet on a deeper level he must be acknowledged the direct and natural heir of Petrarch, Boccaccio, Castiglione, and Montaigne. Like them he delighted more in presenting issues than in espousing systems and held critical awareness, as opposed to doctrinal rectitude, to be the highest possible good. His plays reflect an inquiry into human character entirely in accord with the humanistic emphasis on the dignity of the emotions, and indeed it may be said that his unprecedented use of language as a means of psychological revelation gave striking support to the humanistic contention that language was the heart of culture and the index of the soul. Similarly, Shakespeare's unparalleled realism may be seen as the ultimate embodiment, in poetic terms, of the intense concern for specificity—be it in description, measurement, or imitation—endorsed across the board by humanists from Boccaccio and Salutati on. Shakespearean drama is a treasury of the disputes that frustrated and delighted humanism, including (among many others) action versus contemplation, theory versus practice, *res* versus *verbum,* monarchy versus republic, human dignity versus human depravity, and individualism versus communality. In treating of these polarities, he generally proceeds in the manner of Castiglione and Montaigne, presenting structures of balanced contraries rather than syllogistic endorsements of one side or another. In so doing, he achieves a higher realism, transcending the mere imitation of experience and creating, in all its conflict and fertility, a mirror of mind itself. Since the achievement of such psychological and cultural self-awareness was the primary goal of humanistic inquiry, and since humanists agreed that poetry was an uncommonly effective medium for this achievement, Shakespeare must be acknowledged as a preeminent humanist.

One cannot leave Shakespeare and the phenomenon of English humanism without reference to a highly important aspect of his later drama. Throughout his career, Shakespeare had shown a keen interest in the concept of art, not only as a general idea but also with specific reference to his own identity as dramatist. In two of his final plays, *The Winter's Tale* and *The Tempest,* he developed this concept into dramatic and thematic structures that had strongly doctrinal implications. Major characters in both plays practice a moral artistry—a kind of *humanitas* compounded of awareness, experience, imagination, compassion, and craft—that enables them to beguile and dominate other characters and to achieve enduring justice. This special skill, which is cognate with Shakespeare's own dramatic art, suggests a hypothetical solution to many of

the dilemmas posed in his earlier work. It implies that problems unavailable to political or religious remedy may be solved by creative innovation and that the art by which things are known and expressed may constitute, in and of itself, a valid field of inquiry and an instrument for cultural renewal. In developing this idea of the sovereignty of art, Shakespeare made the final major contribution to a humanistic tradition that will be discussed in the two sections that follow.

HUMANISM AND THE VISUAL ARTS

Humanistic themes and techniques were woven deeply into the development of Italian Renaissance art; conversely, the general theme of "art" was prominent in humanistic discourse. The mutually enriching character of the two disciplines is evident in a variety of areas.

Realism. Humanists paid conscious tribute to realistic techniques in art that had developed independently of humanism. Giotto di Bondone (*c.* 1266–1337), the Florentine painter responsible for the movement away from the Byzantine style and toward ancient Roman technique, was praised by Vasari as "the pupil of Nature." Giotto's own contemporary Boccaccio said of him in the *Decameron* that

> there was nothing in Nature—the mother and ruling force of all created things with her constant revolution of the heavens—that he could not paint with his stylus, pen, or brush or make so similar to its original in Nature that it did not appear to be the original rather than a reproduction. Many times, in fact, in observing things painted by this man, the visual sense of men would err, taking what was painted to be the very thing itself.

Boccaccio, himself a naturalist and a realist, here subtly adopts the painter's achievement as a justification for his own literary style. So Shakespeare, at the end of the Renaissance, praises Giulio Romano (and himself), "who, had he himself eternity and could put breath into his work, would beguile Nature of her custom, so perfectly he is her ape" (*The Winter's Tale*). It should be noted that neither Vasari, Boccaccio, nor Shakespeare endorses realistic style as a *summum bonum:* realism is rather the means for regaining touch with the sovereign creative principle of Nature.

Classicism. Like the humanists, Italian artists of the 15th century saw a profound correlation between classical forms and realistic technique. Classical sculpture and Roman painting were emulated because of their ability to simulate perceived phenomena, while, more abstractly, classical myth offered a unique model for the artistic idealization of human beauty. Alberti, himself a close friend of Donatello and Brunelleschi, codified this humanistic theory of art, using the fundamental principle of mathematics as a link between perceived reality and the ideal. He developed a classically based theory of proportionality between architectural and human form, believing that the ancients sought "to discover the laws by which Nature produced her works so as to transfer them to the works of architecture."

Anthropocentricity and individualism. Humanism and Italian art were similar in giving paramount attention to human experience, both in its everyday immediacy and in its positive or negative extremes. The religious themes that dominated Renaissance art (partly because of generous church patronage) were frequently developed into images of such human richness that, as one contemporary observer noted, the Christian message was submerged. The human-centredness of Renaissance art, moreover, was not just a generalized endorsement of earthly experience. Like the humanists, Italian artists stressed the autonomy and dignity of the individual. High Renaissance art boasted a style of portraiture that was at once humanely appreciative and unsparing of detail. Heroes of culture such as Federico da Montefeltro and Lorenzo de' Medici, neither of whom was a conventionally handsome man, were portrayed realistically, as though a compromise with strict imitation would be an affront to their dignity as individuals. Similarly, artists of the Italian Renaissance were, characteristically, unabashed individualists. The biographies of Giotto, Brunelleschi, Leonardo, and Michelangelo by Giorgio Vasari (1511–74) not only describe artists who were well aware of their unique positions in society and history but also attest to a cultural climate in which, for the first time, the role of art achieved heroic stature. The autobiographical writings of the humanist Alberti, the scientist Gerolamo Cardano (1501–76), and the artist Benvenuto Cellini (1500–71) further attest to the individualism developing both in letters and in the arts; and Montaigne dramatized the analogy between visual mimesis and autobiographical realism when he said, in the preface to his *Essays,* that given the freedom he would have painted himself "*tout entier, et tout nu*" ("totally complete, and totally nude").

Art as philosophy. Italian Renaissance painting, especially in its secular forms, is alive with visually coded expressions of humanistic philosophy. Symbol, structure, posture, and even colour were used to convey silent messages about humanity and nature. Renaissance style was so articulate, and the Renaissance sense of the unity of experience so deeply ingrained, that even architectural structures could be eloquently philosophical. Two features of Federico's palace at Urbino exemplify the profound interrelationship between humanistic principle and Renaissance art. The first feature is architectural. On the ground floor of the palace two private chapels, of roughly the same dimensions, stand side by side. The chapel at the left is a place of Christian worship, while that at the right is dedicated to the pagan Muses. Directly above these chapels is a study, the walls of which are covered with representations (in intarsia) of assorted humanistic heroes: Homer, Plato, Aristotle, Cicero, Virgil, Seneca, Boethius, St. Augustine, Dante, Petrarch, Bessarion, and Federico's revered teacher Vittorino, among others. The message conveyed by the positioning of the three rooms is hard to ignore. Devotion to the opposed principles of Christianity

SCALA/Art Resource, NY



Federico da Montefeltro, duke of Urbino, with his son Guidobaldo, portrait probably by Pedro Berruguete, 15th century. In the National Gallery of the Marches, Urbino, Italy.

Giotto

Biography and autobiography

and earthly (pagan) beauty is rendered possible by a humanistic learning (represented by the study) so generous and appreciative as to comprehend both extremes.

The second feature is iconographic—a portrait of Federico and his son Guidobaldo (probably by Pedro Berruguete) that occupies a central position on the wall of the study. It depicts the Duke, his full coat of armour partly covered by a courtly robe, sitting and reading. The son stands beside his father's chair, gazing out of the picture toward the viewer's left. An abbot's mitre rests on a shelf in the upper left, while the Duke's helmet sits on the floor in the lower right. Here also a typically humanistic message is evident. The Duke's scholarly attitude and curious attire suggest his triple role as warrior, ruler, and humanist. The two main axes of the picture—the line between mitre and helmet and the line between father and son—converge at the book, symbolizing the central role of humanistic learning in reconciling the concerns of church and state and in conveying humanistic virtue from generation to generation. The boy's outward gaze implies the characteristic direction of humanistic learning: into the world of action. The scope and organic wholeness of Federico's humanistic iconography are so striking as to rival great expressions of religious faith. The private heart of his palace concealed, like a genetic code, the principle that had given shape to the edifice and informed the state.

HUMANISM, ART, AND SCIENCE

It is impossible to speak knowledgeably about Renaissance science without first understanding the Renaissance concept of art. The Latin *ars* (inflected as *artis*) was applied indiscriminately to the verbal disciplines, mathematics, music, and science (the "liberal arts"), as well as to painting, sculpture, and architecture; it also could refer to technological expertise, to magic, and to alchemy. Any discipline involving the cultivation of skill and excellence was de facto an art. To the Renaissance, moreover, all arts were "liberal" arts in their capacity to "free" their practitioners to function effectively in specific areas. The art of rhetoric empowered the rhetorician to convince; the art of perspective empowered the painter to create visual illusion; the art of physics empowered the scientist to predict the force and motion of objects. "Art," in effect, was no more or less than articulate power, the technical or intellectual analogy to the political power of the monarch and the divine power of the god. The historical importance of this equation cannot be overestimated. If one concept may be said to have integrated all the varied manifestations of Renaissance culture and given organic unity to the period, it was this definition of art as power. With this definition in mind, one may understand why Renaissance humanists and painters assigned themselves such self-consciously heroic roles: in their artistic ability to delight, to captivate, to convince, they saw themselves as enfranchised directors and remakers of culture. One may also understand why a humanist-artist-scientist like Alberti would have seen no real distinction between the various disciplines he practiced. As profoundly interconnected means of understanding nature and humanity, and as media for effective reform and renewal, these disciplines were all components of an encompassing art. A similar point may be made about Machiavelli, who wrote a book about the "art" of warfare and who used history and logic to develop an art of government, or about the brilliant polymath Paracelsus, who spent his whole career perfecting an art that would comprehend all matter and all spirit. With the equation of art and power in mind, finally, one may understand why a revolutionary scientist like Galileo (1564–1642) put classical and medieval science through a winnowing fan, keeping only such components as allowed for physically reproducible results. Since every Renaissance art aimed for a dominion or conquest, it was completely appropriate that science should leave its previously contemplative role and focus upon the conquest of nature.

The "liberal arts"

Humanism benefited the development of science in a number of more specific ways. Alberti's technological applications of mathematics, and his influential statement that mathematics was the key to all sciences, grew out of his humanistic education at Padua. Vittorino, another student at Padua, went on to make mathematics a central feature of his educational program. Gerolamo Cardano, a scholar of renowned humanistic skills, made major contributions to the development of algebra. In short, the importance of mathematics in humanistic pedagogy and the fact that major humanists like Vittorino and Alberti were also mathematicians may be seen as contributing to the critical role mathematics would play in the rise of modern science. Humanistic philology, moreover, supplied scientists with clean texts and clear Latin translations of the classical works—Plato, Aristotle, Euclid, Archimedes, and even Ptolemy—that furthered their studies. The richness of the classical heritage in science is often underestimated. Galileo, who considered Archimedes his mentor, also prized the dialogues of Plato, in particular the *Meno*. The German philosopher Ernst Cassirer has demonstrated the likelihood that Galileo was fond of the *Meno* because it contained the first statement of the "hypothetical" method, a modus operandi that characterized Galileo's own scientific practice and that would come to be known as one of the chief principles of the New Science. Humanism may also be seen as offering, of itself, methods and attitudes suitable for application in nonhumanistic fields. It might be argued, for example, that the revolutionary social science of Machiavelli and Juan Luis Vives (1492–1540) was due in large measure to their application of humanistic techniques to fields that lay outside the normal purview of humanism. But most of all it was the general spirit of humanism—critical, questing, ebullient, precise, focused on the physical world, and passionate in its quest for results—that fostered the development of the scientific spirit in social studies and natural philosophy.

The classical heritage in science

HUMANISM AND CHRISTIANITY

Though much humanistic activity was specifically Christian in intention, and though the majority of humanists made firm avowals of faith, the relationship between Christianity and humanism is complex and not wholly untroubled. First, humanists from Petrarch onward recognized that the classical (pagan) direction of humanism necessarily constituted, if not a challenge to Christianity, at least a breach in the previous totality of Christian devotion. The Christian truth that had been acknowledged as comprehending all phenomena, earthly or heavenly, now had to coexist with a classical attitude that was overwhelmingly directed toward earthly life. Humanistic efforts to resolve the contradictions implied by these two attitudes were, if one may judge by their variety, never wholly successful. In particular, the extent to which humanistic inquiry led scholars toward the secular realm, and the extent to which humanistic pedagogy concentrated on secular subjects, suggest erosions of the domain of faith. Coluccio Salutati, who urged the young Poggio not to let humanistic enthusiasm take precedence over Christian piety, thereby acknowledged a dualism implicit in the humanistic program and never wholly absent from its historical development.

Implicit dualism

Second, the humanistic philology that meticulously compared ancient sources and "cleaned up" the texts of important Christian writings was a serious challenge to the authority of the church. With new authorities or refined texts in hand, humanists found fault with established commentaries and questioned traditional interpretations. Valla's arraignment of the Donation of Constantine and Bessarion's discovery that the supposed Dionysius the Areopagite (later called Pseudo-Dionysius) had borrowed some of his material from Plato exemplify the uneasy relationship between humanism and Catholic dogma. Third, the independent and broadly critical attitude innate to humanism could not but threaten the unanimity of Christian belief. Intellectual individualism, which has never been popular in any church, put particular stress on a religion that encouraged simple faith and alleged universal authority. Finally, humanism repeatedly fostered the impulse of religious reform. The humanistic emphasis on total authenticity and direct contact with sources had, as its religious correlative, a desire to obliterate the medieval accretions and procedural complexities that stood between the worshiper and his god. The reform-mindedness of

The impulse of reform

such humanists as Petrarch, Boccaccio, Erasmus, and Rabelais was balanced on the religious side by reformers such as Calvin and Melanchthon, who employed humanistic techniques in their own cause. And the reform movement, while it may have modernized and thus preserved Christianity, rang the death knell for a medieval culture whose essential characteristic had been participation in a universal church.

LATER FORTUNES OF HUMANISM

Shakespeare may be seen as the last major interpreter of the humanistic program. Sir Francis Bacon and John Milton, though formidably adept at humanistic techniques, diverged in their major work from the central current of humanism, Bacon toward natural science, Milton toward theology. If Bacon's rationalism may be seen as a link between humanism and the Enlightenment, his strong emphasis on nature (rather than humanity) as subject matter presaged the permanent separation of the sciences from the humanities. In Milton's theocentricity, on the other hand, lay the Christian distrust (going back, perhaps, to Luther) of humanistic secularism. These epochal divergences, moreover, were complemented by a series of rifts and ramifications within the humanistic movement. The split between philosophy and letters was, over future generations, to be compounded by the development of countless discrete specialties within both fields. Philosophers came more and more to define themselves within narrow boundaries. Creative writers and "critics" took up distinct positions and assumed adversarial relationships. The profound loss of coherence in humane letters was furthered by the gradual decline of Latin as the lingua franca of European intellectuals and the consequent separation of national traditions.

By the 19th century, humanism was such a lost art as to have to be reassembled, like a disjointed fossil, by careful historians. Of course there were exceptions. Jonathan Swift (1667–1745) reasserted humanistic values in a broad-based attack on contemporary institutions, and in Gottfried Wilhelm Leibniz (1646–1716) can be found the serious intention and multifarious curiosity that characterized humanism at its best. Strong humanistic motives may be found in Germany at the turn of the 19th century, particularly in the work of Gotthold Ephraim Lessing (1729–81), Friedrich von Schiller (1759–1805), and Georg Wilhelm Friedrich Hegel (1770–1831); while Johann Wolfgang von Goethe (1749–1832) was perhaps the last individual whose breadth of achievement and sense of the unity of experience lived up to the ideal established by Alberti.

Goethe

More recently, the mode of inquiry and interpretation developed by the political philosopher Leo Strauss (1899–1973) showed strong signs of the humanistic spirit. But in general the traces of the original program have been scattered. To the modern mind, a "humanist" is a university scholar, walled off from the interdisciplinary scope of the original humanistic program and immune to the active experience that was its basis and its goal. This decline is easy enough to explain. Had there been nothing else, one external factor would have made the cultivation of *humanitas,* as originally practiced, more and more difficult from the beginning of the 16th century on. The proliferation of published work in all fields, and the creation of many new fields, made increasingly impracticable the development of the comprehensive learning and awareness that were central to the original program. In 1500 the major texts constituting a humanistic education, though numerous, could still be counted; by 1900 they were legion, and people had long ceased agreeing about exactly which ones they were. But problems implicit in the movement were equally responsible for its demise. The characteristic emphases on rhetoric and philology, which gave the humanistic movement vitality and made it available to countless students of moderate gifts, also betokened its impermanence. Weak in dialectic or any other comprehensively analytic method, the movement had no instrument for self-examination, no medium for self-renewal. By the same token, neither had humanism any valid means of defense against the attackers—scientists, fundamentalists, materialists, and others—who camped in ever larger numbers on its borders. Lacking an integral method, finally, humanism in effect lacked a centre and became prey to an endless series of ramifications. While eloquent humanists rambled through Europe and spread the word about the classics, the method that might have unified their efforts lay, available but unheeded, in texts of Plato and Aristotle. Given this core of rigorous analysis, humanism might (all other challenges notwithstanding) have retained its basic character for centuries. But ironically it might also have failed to attract followers.

The lack of method

CONCLUSION

Though lacking permanence itself, humanism in large measure established the climate and provided the medium for the rise of modern thought. An impressive variety of major developments in literature, philosophy, art, religion, social science, and even natural science had their basis in humanism or were significantly nourished by it. Important spokesmen in all fields regularly made use of humanistic eloquence to further their causes. More generally, the so-called modern awareness—that sense of alienation and freedom applied both to the individual and to the race—derives ultimately, for better or worse, from humanistic sources. But with humanism, as with every other historical subject, one should beware lest valid concern about changes, crises, sources, and influences obscure the even more important issues of human continuity and human value. Whatever its weaknesses and inner conflicts, the humanistic movement was heroic in its breadth and energy, remarkable in its aspirations. For human development in all fields, it created a context of seldom-equaled fertility. Its characteristic modalities of thought, speech, and image lent themselves to the promptings of genius and became the media for enduring achievement. Its moral program formed the basis for lives that are remembered with admiration.

BIBLIOGRAPHY

General treatments: The three general studies most helpful in approaching the phenomenon of humanism are EUGENIO GARIN, *Italian Humanism: Philosophy and Civic Life in the Renaissance,* trans. from the Italian (1965, reprinted 1975); PAUL OSKAR KRISTELLER, *Renaissance Thought and Its Sources* (1979); and CHARLES TRINKAUS, *The Scope of Renaissance Humanism* (1983). Garin's book is probably the most unified and incisive treatment of Italian humanism yet produced, while Kristeller and Trinkaus offer extremely well-documented analyses of major issues in the history and historiography of humanism. Other valuable readings include HANS BARON, *The Crisis of the Early Italian Renaissance: Civic Humanism and Republican Liberty in an Age of Classicism and Tyranny,* rev. ed. (1966); QUIRINUS BREEN, *Christianity and Humanism: Studies in the History of Ideas* (1968); JACOB BURCKHARDT, *The Civilization of the Period of the Renaissance in Italy: An Essay* (1878; originally published in German, 1860), available in later English-language editions; DOUGLAS BUSH, *The Renaissance and English Humanism* (1939, reprinted 1972); ERNST CASSIRER, *The Individual and the Cosmos in Renaissance Philosophy* (1964, reprinted 1972; originally published in German with appendices, 1927); JACK D'AMICO, *Knowledge and Power in the Renaissance* (1977); MYRON P. GILMORE, *The World of Humanism, 1453–1517* (1952; reprinted 1983); DENYS HAY, *The Italian Renaissance in Its Historical Background,* 2nd ed. (1977); PAUL OSKAR KRISTELLER, *Renaissance Concepts of Man, and Other Essays* (1972); EDWARD P. MAHONEY (ed.), *Philosophy and Humanism* (1976); ROBERT MANDROU, *From Humanism to Science, 1480 to 1700* (1979; originally published in French, 1973); HEIKO A. OBERMAN and THOMAS A. BRADY, JR. (eds.), *Itinerarium Italicum: The Profile of the Italian Renaissance in the Mirror of Its European Transformations* (1975); CHARLES B. SCHMITT, *Studies in Renaissance Philosophy and Science* (1981); JOHN ADDINGTON SYMONDS, *Renaissance in Italy,* 7 vol. (1875–86, reprinted 1971–72); GUISSEPPE TOFFANIN, *History of Humanism* (1954; originally published in Italian, 1933); BERTHOLD L. ULLMAN, *Studies in the Italian Renaissance,* 2nd ed. (1975); and ROBERTO WEISS, *The Spread of Italian Humanism* (1964). For classical and medieval backgrounds, see ERNST ROBERT CURTIUS, *European Literature and the Latin Middle Ages* (1973; originally published in German, 1948); MOSES HADAS, *Humanism: The Greek Ideal and Its Survival* (1960, reprinted 1972); and WERNER JAEGER, *Paideia: The Ideals of Greek Culture,* 2nd ed., 3 vol. (1965; originally published in German, 1934).

Specific topics: T.W. BALDWIN, *William Shakspere's Small*

Latine and Lesse Greeke (1944, reprinted 1966); HANS BARON, *From Petrarch to Leonardo Bruni: Studies in Humanistic and Political Literature* (1968); GENE BRUCKER, *Renaissance Florence* (1969, reprinted 1983); ERNST CASSIRER, "Galileo's Platonism," in M.F. ASHLEY MONTAGU (ed.), *Studies and Essays in the History of Science and Learning,* pp. 277–297 (1946; reprinted 1975); JOAN GADOL, *Leon Battista Alberti* (1969); EUGENIO GARIN, *Science and Civic Life in the Italian Renaissance* (1969, reissued 1978; originally published in Italian, 1965; 4th Italian ed., 1980); PAUL OSKAR KRISTELLER and PHILIP P. WIENER (eds.), *Renaissance Essays* (1968); LAURO MARTINES, *The Social World of the Florentine Humanists, 1390–1460* (1963); JAMES J. MURPHY (ed.), *Renaissance Eloquence: Studies in the Theory and Practice of Renaissance Rhetoric* (1983); IRWIN PANOFSKY, *Renaissance and Renascences in Western Art,* 2 vol. (1960, reissued in 1 vol, 1969); J.H. PLUMB (ed.), *Renaissance Profiles* (1965); PASQUALE ROTONDI, *The Ducal Palace of Urbino: Its Architecture and Decoration* (1969; originally published in Italian in 2 vol., 1950–51); CHARLES B. SCHMITT, *Aristotle and the Renaissance* (1983); CHARLES TRINKAUS, *In Our Image and Likeness: Humanity and Divinity in Italian Humanist Thought,* 2 vol. (1970), and *The Poet as Philosopher: Petrarch and the Formation of Renaissance Consciousness* (1979); BERTHOLD L. ULLMAN, *The Humanism of Coluccio Salutati* (1963); WILLIAM HARRISON WOODWARD, *Vittorino da Feltre and Other Humanist Educators* (1897, reissued 1970), and *Studies in Education During the Age of the Renaissance, 1400–1600* (1906, reissued 1967); and G.F. YOUNG, *The Medici,* 2 vol. (1909, reissued in 1 vol., 1933).

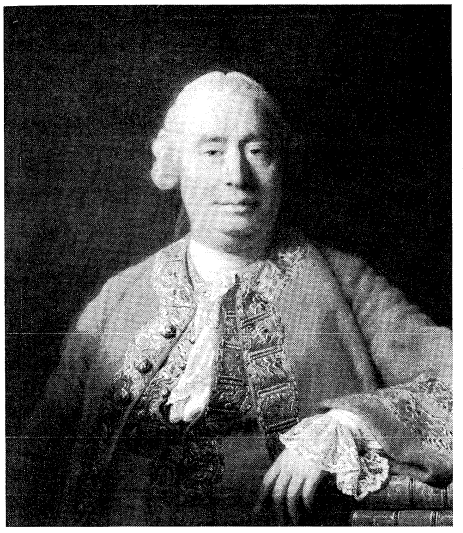
Works of the humanists: Works by the later humanists (*c.* 1500 and after) and the English poet-humanists mentioned in the article, including Castiglione, Cellini, Elyot, Erasmus, Jonson, Machiavelli, Montaigne, More, Pico della Mirandola, Rabelais, Shakespeare, Sidney, Spenser, Tasso, and Vasari, are readily available in many modern English editions. For the writings of the earlier humanists, see LEON BATTISTA ALBERTI, *The Family in Renaissance Florence,* trans. by RENÉE NEU WATKINS (1969); GIOVANNI BOCCACCIO, *The Decameron,* trans. by MARK MUSA and PETER BONDANELLA (1982), and *Boccaccio on Poetry,* 2nd ed., ed. and trans. by CHARLES G. OSGOOD (1956, reprinted 1978), a translation of the preface and books 14 and 15 of his *De genealogia deorum gentilium;* biographies of Dante by Boccaccio and Leonardo Bruni in *The Earliest Lives of Dante,* trans. by JAMES ROBINSON SMITH (1901, reprinted 1976); letters by Poggio Bracciolini in *Two Renaissance Book Hunters,* trans. by PHYLLIS WALTER GOODHART GORDAN (1974); and works by Petrarch, including *The Life of Solitude,* trans. by JACOB ZEITLIN (1924, reprinted 1978); "On His Own Ignorance and That of Many Others," trans. by HANS NACHOD in ERNST CASSIRER, PAUL OSKAR KRISTELLER, and JOHN HERMAN RANDALL, JR. (eds.), *The Renaissance Philosophy of Man* (1948, reprinted 1971); and *Petrarch's Secret; or, The Soul's Conflict with Passion,* trans. by WILLIAM H. DRAPER (1911, reprinted 1978). WILLIAM HARRISON WOODWARD, *op. cit.,* contains valuable translations of works by Vergerio, Bruni, Aeneas Sylvius Piccolomini (Pius II), and Battista Guarino. Sixteenth-century writings that reflect the breadth and vitality of humanistic attitudes are JUAN LUIS VIVES, *On Education,* trans. by FOSTER WATSON (1913, reprinted 1971); GIROLAMO CARDANO, *The Book of My Life,* trans. by JEAN STONER (1930, reprinted 1962); TORQUATO TASSO, *Tasso's Dialogues,* trans. by CARNES LORD and DAIN A. TRAFTON (1982); and PARACELSUS, *Selected Writings,* trans. by NORBERT GUTERMAN, 2nd rev. ed. (1958; reissued with a new bibliography, 1969; originally published in German, 1942, ed. by JOLANDE JACOBI).

(Ro.G.)

Hume

David Hume was an 18th-century Scottish empiricist philosopher, historian, economist, and essayist, who conceived of philosophy as the inductive, experimental science of human nature. Taking the scientific method of the English physicist Sir Isaac Newton as his model and building on the epistemology of the English philosopher John Locke, Hume tried to describe how the mind works in acquiring what is called knowledge. He concluded that no theory of reality is possible; there can be no knowledge of anything beyond experience. Despite the enduring impact of his theory of knowledge, Hume seems to have considered himself chiefly as a moralist.

By courtesy of the Scottish National Portrait Gallery



Hume, oil painting by Allan Ramsay, 1766. In the Scottish National Portrait Gallery, Edinburgh.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* Part Ten, Division V, especially Section 51.

Early life and works. Hume was the younger son of Joseph Hume, the modestly circumstanced laird, or lord, of Ninewells, a small estate adjoining the village of Chirnside, about nine miles distant from Berwick-upon-Tweed on the Scottish side of the border. David's mother, Catherine, a daughter of Sir David Falconer, president of the Scottish court of session, was in Edinburgh when he was born, on May 7 (April 26, old style), 1711. In his third year his father died. He entered Edinburgh University when he was about 12 years old and left it at 14 or 15, as was then usual. Pressed a little later to study law (in the family tradition on both sides), he found it distasteful and instead read voraciously in the wider sphere of letters. Because of the intensity and excitement of his intellectual discovery, he had a nervous breakdown in 1729, from which it took him a few years to recover.

Writing the *Treatise*

In 1734, after trying his hand in a merchant's office in Bristol, he came to the turning point of his life and retired to France for three years. Most of this time he spent at La Flèche on the Loire, in the old Anjou, studying and writing *A Treatise of Human Nature.* The *Treatise* was Hume's attempt to formulate a full-fledged philosophical system. It is divided into three books: book I, on understanding, aims at explaining man's process of knowing, describing in order the origin of ideas, the ideas of space and time, causality, and the testimony of the senses; book II, on the "passions" of man, gives an elaborate psychological machinery to explain the affective, or emotional, order in man and assigns a subordinate role to reason in this mechanism; book III, on morals, describes moral goodness in terms of "feelings" of approval or disapproval that a person has when he considers human behaviour in the light of the agreeable or disagreeable consequences either to himself or to others. Although the *Treatise* is Hume's most thorough exposition of his thought, at the end of his life he vehemently repudiated it as juvenile, avowing that only his later writings presented his considered views. The *Treatise* is not well constructed, in parts oversubtle, confusing because of ambiguity in important terms (especially "reason"), and marred by willful extravagance of statement and rather theatrical personal avowals. For these reasons his mature condemnation of it was perhaps not entirely misplaced. Book I, nevertheless, has been more read in academic circles than any other of his writings.

Returning to England in 1737, he set about publishing the *Treatise.* Books I and II were published in two volumes in 1739; book III appeared the following year. The poor reception of this, his first and very ambitious work, depressed him; but his next venture, *Essays, Moral and Political* (1741–42), won some success. Perhaps encouraged by this, he became a candidate for the chair of moral philosophy at Edinburgh in 1744. Objectors alleged heresy and even atheism, pointing to the *Treatise* as evidence. Unsuccessful, Hume left the city, where he had been living since 1740, and began a period of wandering: a sorry year near St. Albans as tutor to the mad marquess of Annandale (1745–46); a few months as secretary to Gen. James St. Clair (a member of a prominent Scottish family), with whom he saw military action during an abortive expedition to Brittany (1746); a little tarrying in London and at Ninewells; and then some further months with General St. Clair on an embassy to the courts of Vienna and Turin (1748–49).

Mature works. During his years of wandering Hume was earning the money that he needed to gain leisure for his studies. Some fruits of these studies had already appeared before the end of his travels, viz., a further *Three Essays, Moral and Political* (1748) and *Philosophical Essays Concerning Human Understanding* (1748). The latter is a rewriting of book I of the *Treatise* (with the addition of his essay "On Miracles," which became notorious for its denial that a miracle can be proved by any amount or kind of evidence); it is better known as *An Enquiry Concerning Human Understanding,* the title Hume gave to it in a revision of 1758. The *Enquiry Concerning the Principles of Morals* (1751) was a rewriting of book III of the *Treatise.* It was in these works that Hume expressed his mature thought.

Theory of knowledge

An Enquiry Concerning Human Understanding is an attempt to define the principles of human knowledge. It poses in logical form significant questions about the nature of reasoning in regard to matters of fact and experience, and it answers them by recourse to the principle of association. The basis of his exposition is a twofold classification of objects of awareness. In the first place, all such objects are either "impressions," data of sensation or of internal consciousness, or "ideas," derived from such data by compounding, transposing, augmenting, or diminishing. That is to say, the mind does not create any ideas but derives them from impressions. From this Hume develops a theory of meaning. A word that does not stand directly for an impression has meaning only if it brings before the mind an object that can be gathered from an impression by one of the mental processes mentioned. In the second place, there are two approaches to construing meaning, an analytical one, which concentrates on the "relations of ideas," and an empirical one, which focuses on "matters of fact." Ideas can be held before the mind simply as meanings, and their logical relations to one another can then be detected by rational inspection. The idea of a plane triangle, for example, entails the equality of its internal

angles to two right angles, and the idea of motion entails the ideas of space and time, irrespective of whether there really are such things as triangles and motion. Only on this level of mere meanings, Hume asserts, is there room for demonstrative knowledge. Matters of fact, on the other hand, come before the mind merely as they are, revealing no logical relations; their properties and connections must be accepted as they are given. That primroses are yellow, that lead is heavy, and that fire burns things are facts, each shut up in itself, logically barren. Each, so far as reason is concerned, could be different: the contradictory of every matter of fact is conceivable. Therefore, any demonstrative science of fact is impossible.

Hume's doctrine of causality and belief

From this basis Hume develops his doctrine about causality. The idea of causality is alleged to assert a necessary connection among matters of fact. From what impression, then, is it derived? Hume states that no causal relation among the data of the senses can be observed, for, when a person regards any events as causally connected, all that he does and can observe is that they frequently and uniformly go together. In this sort of togetherness it is a fact that the impression or idea of the one event brings with it the idea of the other. A habitual association is set up in the mind; and, as in other forms of habit, so in this one, the working of the association is felt as compulsion. This feeling, Hume concludes, is the only discoverable impressional source of the idea of causality.

Hume then considers the process of causal inference, and in so doing he introduces the concept of belief. When a person sees a glass fall, he not only thinks of its breaking but expects and believes that it will break; or, starting from an effect, when he sees the ground to be generally wet, he not only thinks of rain but believes that there has been rain. Thus belief is a significant component in the process of causal inference. Hume then proceeds to investigate the nature of belief, claiming that he was the first to do so. He uses this term in the narrow sense of belief regarding matters of fact. He defines belief as a sort of liveliness or vividness that accompanies the perception of an idea. A belief is more than an idea; it is a vivid or lively idea. This vividness is originally possessed by some of the objects of awareness, by impressions and the simple memory images of them. By association it comes to belong to certain ideas as well. In the process of causal inference, then, an observer passes from an impression to an idea regularly associated with it. In the process the aspect of liveliness proper to the impression infects the idea, Hume asserts. And it is this aspect of liveliness that Hume defines as the essence of belief.

Hume does not claim to prove that the propositions, (1) that events themselves are causally related and (2) that they will be related in the future in the same ways as they were in the past, are false. He firmly believed both of these propositions and insisted that everybody else believed them, will continue to believe them, and must continue to believe them in order to survive. They are natural beliefs, inextinguishable propensities of human nature, madness apart. What Hume claims to prove is that natural beliefs are not obtained and cannot be demonstrated either by empirical observation or by reason, whether intuitive or inferential. Reflection shows that there is no evidence for them and shows also both that we are bound to believe them and that it is sensible or sane to do so. This is Hume's skepticism: it is an affirmation of that tension, a denial not of belief but of certainty.

Theory of morals and historical writing

The *Enquiry Concerning the Principles of Morals* is a refinement of Hume's thinking on morality, in which he views sympathy as the fact of human nature lying at the basis of all social life and personal happiness. Defining morality as those qualities that are approved (1) in whomsoever they happen to be and (2) by virtually everybody, he sets himself to discover the broadest grounds of the approvals. He finds them, as he found the grounds of belief, in "feelings," not in "knowings." Moral decisions are grounded in moral sentiment. Qualities are valued either for their utility or for their agreeableness, in each case either to their owners or to others. Hume's moral system aims at the happiness of others (without any such formula as "the greatest happiness of the greatest number") and at the happiness of self. But regard for others accounts for the greater part of morality. His emphasis is on altruism: the moral sentiments that he claims to find in human beings, he traces, for the most part, to a sentiment for and a sympathy with one's fellows. It is human nature, he holds, to laugh with the laughing and to grieve with the grieved and to seek the good of others as well as one's own. Two years after the *Enquiry* was published, Hume confessed, "I have a partiality for that work"; and at the end of his life he judged it "of all my writings incomparably the best." Such statements, along with other indications in his later writings, make it possible to suspect that he regarded his moral doctrine as his major work. He here writes as a man having the same commitment to duty as his fellows. The traditional view that he was a detached scoffer is deeply wrong: he was skeptical not of morality but of much theorizing about it.

Following the publication of these works, Hume spent several years (1751–63) in Edinburgh, with two breaks in London. An attempt was made to get him appointed as successor to Adam Smith, the Scottish economist (later to be his close friend), in the chair of logic at Glasgow, but the rumour of atheism prevailed again. In 1752, however, Hume was made keeper of the Advocates' Library at Edinburgh. There, "master of 30,000 volumes," he could indulge a desire of some years to turn to historical writing. His *History of England,* extending from Caesar's invasion to 1688, came out in six quarto volumes between 1754 and 1762, preceded by *Political Discourses* (1752). His recent writings had begun to make him known, but these two brought him fame, abroad as well as at home. He also wrote *Four Dissertations* (1757), which he regarded as a trifle, although it included a rewriting of book II of the *Treatise* (completing his purged restatement of this work) and a brilliant study of "the natural history of religion." In 1762 James Boswell, the biographer of Samuel Johnson, called Hume "the greatest writer in Britain," and the Roman Catholic Church, in 1761, paid him the attention of putting all his writings on the *Index,* its list of forbidden books.

Hume's reception in Paris

The most colourful episode of his life ensued: in 1763 he left England to become secretary to the British embassy in Paris under the Earl of Hertford. The society of Paris accepted him, despite his ungainly figure and gauche manner. He was honoured as eminent in breadth of learning, in acuteness of thought, and in elegance of pen and was taken to heart for his simple goodness and cheerfulness. The salons threw open their doors to him, and he was warmly welcomed by all. For four months in 1765 he acted as chargé d'affaires at the embassy. When he returned to London at the beginning of 1766 (to become, a year later, undersecretary of state), he brought Jean-Jacques Rousseau, the Swiss philosopher connected with the *Encyclopédie* of Diderot and d'Alembert, with him and found him a refuge from persecution in a country house at Wootton in Staffordshire. This tormented genius suspected a plot, took secret flight back to France, and spread a report of Hume's bad faith. Hume was partly stung and partly persuaded into publishing the relevant correspondence between them with a connecting narrative (*A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau,* 1766).

In 1769, somewhat tired of public life and of England too, he again established a residence in his beloved Edinburgh, deeply enjoying the company—at once intellectual and convivial—of friends old and new (he never married), as well as revising the text of his writings. He issued five further editions of his *History* between 1762 and 1773 as well as eight editions of his collected writings (omitting the *Treatise, History,* and ephemera) under the title *Essays and Treatises* between 1753 and 1772, besides preparing the final edition of this collection, which appeared posthumously (1777), and *Dialogues Concerning Natural Religion,* held back under pressure from friends and not published until 1779. His curiously detached autobiography, *The Life of David Hume, Esquire, Written by Himself* (1777; the title is his own), is dated April 18, 1776. After a long illness he died in his Edinburgh house on Aug. 25, 1776, and was buried on Calton Hill.

Adam Smith, his literary executor, added to the *Life* a letter that concludes with his judgment on his friend as "approaching as nearly to the idea of a perfectly wise and virtuous man as perhaps the nature of human frailty will permit." His distinguished friends, with ministers of religion among them, certainly admired and loved him, and there were younger men indebted either to his influence or to his pocket. The mob had heard only that he was an atheist and simply wondered how such an ogre would manage his dying. Yet Boswell has recounted, in a passage in his *Private Papers,* that, when he visited Hume in his last illness, the philosopher put up a lively, cheerful defense of his disbelief in immortality.

Significance and influence. That Hume was one of the major figures of his century can hardly be doubted. So his contemporaries thought, and his achievement, as seen in historical perspective, confirms that judgment, though with a shift of emphasis. Some of the reasons for the assessment may be given under four heads:

As a writer. Hume's style was praised in his lifetime and has often been praised since. It exemplifies the classical standards of his day. It lacks individuality and colour, for he was always proudly on guard against his emotions. The touch is light, except on slight subjects, where it is rather heavy. Yet in his philosophical works he gives an unsought pleasure. Here his detachment, levelness (all on one plane), smoothness, and daylight clearness are proper merits. It is as one of the best writers of scientific prose in English that he stands in the history of style.

As a historian. Library catalogs still list Hume as "Hume, David, the Historian." Between his death and 1894, there were at least 50 editions of his *History;* and an abridgment, *The Student's Hume* (1859; often reprinted), remained in common use for 50 years. Though now outdated, Hume's *History* must be regarded as an event of cultural importance. In its own day, moreover, it was an innovation, soaring high above its very few predecessors. It was fuller and set a higher standard of impartiality. His *History of England* not only traced the deeds of kings and statesmen but also displayed the intellectual interests of the educated citizens, as may be seen, for instance, in the pages on literature and science under the Commonwealth at the end of chapter 3 and under James II at the end of chapter 2. It was unprecedentedly readable, in structure as well as in phrasing. Persons and events were woven into causal patterns that furnished a narrative with the goals and resting points of recurrent climaxes. That was to be the plan of future history books for the general reader.

As an economist. Hume steps forward as an economist in the *Political Discourses* incorporated in *Essays and Treatises* as part 2 of *Essays Moral and Political.* How far he influenced his friend Adam Smith, 12 years his junior, remains uncertain: they had broadly similar principles, and both had the excellent habit of illustrating and supporting these from history. He did not formulate a complete system of economic theory, as did Adam Smith in his *Wealth of Nations,* but Hume introduced several of the new ideas around which the "classical economics" of the 18th century was built. His level of insight can be gathered from his main contentions: that wealth consists not of money but of commodities; that the amount of money in circulation should be kept related to the amount of goods in the market (two points made by Berkeley); that a low rate of interest is a symptom not of superabundance of money but of booming trade; that no nation can go on exporting only for bullion; that each nation has special advantages of raw materials, climate, and skill, so that a free interchange of products (with some exceptions) is mutually beneficial; and that poor nations impoverish the rest just because they do not produce enough to be able to take much part in that exchange. He welcomed advance beyond an agricultural to an industrial economy as a precondition of any but the barer forms of civilization.

Main contentions in economics

As a philosopher. Hume conceived of philosophy as the inductive science of human nature, and he concluded that man is more a creature of sensitive and practical sentiment than of reason. On the Continent he is seen as one of the few British classical philosophers. For some Germans his importance lies in the fact that Immanuel Kant conceived his critical philosophy in direct reaction to Hume. Hume was one of the influences that led Auguste Comte, the 19th-century French mathematician and sociologist, to positivism. In Britain his positive influence is seen in Jeremy Bentham, the early 19th-century jurist and philosopher, who was moved to utilitarianism (the moral theory that right conduct should be determined by the usefulness of its consequences) by book III of the *Treatise,* and more extensively in John Stuart Mill, the philosopher and economist who lived later in the 19th century.

In throwing doubt on the assumption of a necessary link between cause and effect, Hume was the first philosopher of the postmedieval world to reformulate the skepticism of the ancients. His reformulation, moreover, was carried out in a new and compelling way. Although Hume admired Newton, Hume's subtle undermining of causality called in question the philosophical basis of Newton's science as a way of looking at the world, inasmuch as this rested on the identification of a few fundamental causal laws that govern the universe. As a result the positivists of the 19th century were obliged to wrestle with Hume's questioning of causality if they were to succeed in their aim of making science the central framework of human thought. In the 20th century it was Hume's naturalism rather than his skepticism that attracted attention, chiefly among analytic philosophers. Hume's naturalism lies in his belief that philosophical justification could only be rooted in regularities of the natural world. The attraction of this for analytic philosophers was that it seemed to provide a solution to the problems arising from the skeptical tradition that Hume himself, in his other philosophical role, had done so much to reinvigorate.

MAJOR WORKS

PHILOSOPHY AND RELIGION: *A Treatise of Human Nature: Being an Attempt to Introduce the Experimental Method of Reasoning into Moral Subjects* (1739–40); *An Abstract of a Book Lately Published: Entitled, A Treatise of Human Nature, etc., Wherein the Chief Argument of That Book Is Farther Illustrated and Explained* (1740); *Philosophical Essays Concerning Human Understanding* (1748; many later editions entitled *An Enquiry Concerning Human Understanding*); *Four Dissertations* (1757); *Dialogues Concerning Natural Religion* (1779); *A Letter from a Gentleman to His Friend Containing Some Observations on Religion and Morality* (1745).

POLITICS AND MORALS: *Essays, Moral and Political* (1741–42); *An Enquiry Concerning the Principles of Morals* (1751); *Political Discourses* (1752).

HISTORY: *The History of Great Britain* (1754–57); *The History of England Under the House of Tudor* (1759); *The History of England from the Invasion of Julius Caesar to the Accession of Henry VII* (1762).

OTHER WORKS: *A Concise and Genuine Account of the Dispute Between Mr. Hume and Mr. Rousseau* (1766); *The Life of David Hume, Esquire, Written by Himself* (1777).

Recommended modern editions of separate works by Hume include: *A Treatise of Human Nature,* ed. by L.A. Selby-Bigge, 2nd ed. rev. by P.H. Nidditch (1980); *Enquiries Concerning Human Understanding and Concerning the Principles of Morals,* ed. by L.A. Selby-Bigge, 3rd ed. rev. by P.H. Nidditch (1975); *Dialogues Concerning Natural Religion,* ed. by Norman Kemp Smith (1935, reissued 1981); *The Natural History of Religion,* ed. by A. Wayne Colver (1976); and *The History of Great Britain: The Reigns of James I and Charles I,* ed. by Duncan Forbes (1970). The best collected edition is *The Philosophical Works of David Hume,* ed. by T.H. Green and T.H. Grose, new ed., 4 vol. (1882–86, reprinted 1964).

BIBLIOGRAPHY

Biographies: The standard biography is ERNEST C. MOSSNER, *The Life of David Hume,* 2nd rev. ed. (1980); while JOHN HILL BURTON, *Life and Correspondence of David Hume,* 2 vol. (1846, reprinted 1983), is not entirely superseded by more recent scholarship. Shorter biographical studies include J.Y.T. GREIG, *David Hume* (1931, reprinted 1983); and F.H. HEINEMANN, *David Hume, the Man and His Science of Man* (1940). J.Y.T. GREIG (ed.), *The Letters of David Hume,* 2 vol. (1932, reprinted 1983), is supplemented by RAYMOND KLIBANSKY and ERNEST C. MOSSNER (eds.), *New Letters of David Hume* (1954, reprinted 1983).

Commentaries: An up-to-date introduction to Hume's philosophical ideas is provided by A.J. AYER, *Hume* (1980). Other useful introductory works are D.G.C. MACNABB, *David Hume: His Theory of Knowledge and Morality,* 2nd ed. (1966); BARRY STROUD, *Hume* (1977); and V.C. CHAPPELL (ed.), *The Philosophy*

of David Hume (1963), which contains excerpts from Hume's works. More specialized commentaries on Hume's epistemological theory include H.H. PRICE, *Hume's Theory of the External World* (1940, reprinted 1981); CONSTANCE MAUND, *Hume's Theory of Knowledge* (1937, reprinted 1973); and NORMAN KEMP SMITH, *The Philosophy of David Hume* (1941, reprinted 1983). Commentaries that concentrate on Hume's ethical thought are R. DAVID BROILES, *The Moral Philosophy of David Hume,* 2nd ed. (1969); J.L. MACKIE, *Hume's Moral Theory* (1980); ANTONY FLEW, *Hume's Philosophy of Belief* (1961); RACHEL M. KYDD, *Reason and Conduct in Hume's Treatise* (1946, reissued 1964); and JONATHAN HARRISON, *Hume's Moral Epistemology* (1976), and *Hume's Theory of Justice* (1981, reprinted 1983). ALASDAIR MACINTYRE (ed.), *Hume's Ethical Writings* (1965, reprinted 1979), contains excerpts that are well chosen from various works.

Hume's political theory is examined in DUNCAN FORBES, *Hume's Philosophical Politics* (1975, reprinted 1985); DAVID MILLER, *Philosophy and Ideology in Hume's Political Thought* (1981, reprinted 1984); G.K. VLACHOS, *Essai sur la politique de Hume* (1955); and SHIRLEY ROBIN LETWIN, *The Pursuit of Certainty: David Hume, Jeremy Bentham, John Stuart Mill, Beatrice Webb* (1965). Excerpts from Hume's writings on politics, together with a commentary, appear in FREDERICK WATKINS (ed.), *Theory of Politics* (1951).

RICHARD WOLLHEIM (ed.), *Hume on Religion* (1963, reissued 1964), contains Hume's most important writings on the subject. More extensive commentaries are provided by J.C.A. GASKIN, *Hume's Philosophy of Religion* (1978); and ANDRÉ-LOUIS LEROY, *La Critique et la religion chez David Hume* (1930).

EUGENE ROTWEIN (ed.), *Writings on Economics* (1955, reissued 1972), a collection of Hume's essays, is a valuable introduction to a somewhat neglected aspect of Hume's social theory. Commentaries that attempt a more synoptic and comprehensive appraisal of Hume's thought include JOHN PASSMORE, *Hume's Intentions,* 3rd ed. (1980); CHARLES W. HENDEL, *Studies in the Philosophy of David Hume* (1925, reissued 1983); ANDRÉ-LOUIS LEROY, *David Hume* (1953, reprinted 1979); JOHN LAIRD, *Hume's Philosophy of Human Nature* (1932, reprinted 1983); A.P. CAVENDISH (pseudonym A.H. BASSON), *David Hume* (1958, reprinted 1981); FARHANG ZABEEH, *Hume, Precursor of Modern Empiricism,* 2nd rev. ed. (1973); and JAMES NOXON, *Hume's Philosophical Development* (1973, reprinted 1975).

Bibliographies: T.E. JESSOP, *A Bibliography of David Hume and of Scottish Philosophy from Frances Hutchinson to Lord Balfour* (1938, reprinted 1983); and ROLAND HALL, *Fifty Years of Hume Scholarship* (1978), and *A Hume Bibliography, from 1930* (1971). WILLIAM B. TODD (ed.), *Hume and the Enlightenment* (1974), contains a bibliographical essay. *Hume Studies* (semiannual) updates bibliographical information.

(T.E.Je./M.C.)

Humour and Wit

In all its many-splendoured varieties, humour can be simply defined as a type of stimulation that tends to elicit the laughter reflex. Spontaneous laughter is a motor reflex produced by the coordinated contraction of 15 facial muscles in a stereotyped pattern and accompanied by altered breathing. Electrical stimulation of the main lifting muscle of the upper lip, the zygomatic major, with currents of varying intensity produces facial expressions ranging from the faint smile through the broad grin to the contortions typical of explosive laughter.

The laughter and smile of civilized man is, of course, often of a conventional kind, in which voluntary intent substitutes for, or interferes with, spontaneous reflex activity; this article is concerned, however, only with the latter. Once laughter is realized to be a humble reflex, several paradoxes must be faced. Motor reflexes, such as the contraction of the pupil of the eye in dazzling light, are simple responses to simple stimuli whose value to survival is obvious. But the involuntary contraction of 15 facial muscles, associated with certain irrepressible noises, strikes one as an activity without any utilitarian value, quite unrelated to the struggle for survival. Laughter is a reflex but unique in that it has no apparent biological purpose. One might call it a luxury reflex. Its only function seems to be to provide relief from tension.

The second related paradox is a striking discrepancy between the nature of the stimulus and that of the response in humorous transactions. When a blow beneath the kneecap causes an automatic upward kick, both "stimulus" and "response" function on the same primitive physiological level, without requiring the intervention of the higher mental functions. But that such a complex mental activity as reading a comic story should cause a specific reflex contraction of the facial muscles is a phenomenon that has puzzled philosophers since Plato. There is no clear-cut, predictable response that would tell a lecturer whether he has succeeded in convincing his listeners; but when he is telling a joke, laughter serves as an experimental test. *Humour is the only form of communication in which a stimulus on a high level of complexity produces a stereotyped, predictable response on the physiological reflex level.* Thus the response can be used as an indicator for the presence of the elusive quality that is called humour—as the click of the Geiger counter is used to indicate the presence of radioactivity. Such a procedure is not possible in any other form of art; and since the step from the sublime to the ridiculous is reversible, the study of humour provides clues for the study of creativity in general.

This article deals with the changing concepts and practice of humour from the time of Aristotle to the influence of the mass media in the contemporary world.

The article is divided into the following sections:

The logic of laughter 682
Laughter and emotion 683
Verbal humour 684
Situational humour 685
Styles and techniques in humour 686
Relations to art and science 686
The humanization of humour 687
  Non-Western styles
  Humour in the contemporary world
Bibliography 687

THE LOGIC OF LAUGHTER

The range of laughter-provoking experiences is enormous, from physical tickling to mental titillations of the most varied kinds. There is unity in this variety, however, a common denominator of a specific and specifiable pattern that reflects the "logic" or "grammar" of humour, as it were. A few examples will help to unravel that pattern.

1. A masochist is a person who likes a cold shower in the morning so he takes a hot one.
2. An English lady, on being asked by a friend what she thought of her departed husband's whereabouts: "Well, I suppose the poor soul is enjoying eternal bliss, but I wish you wouldn't talk about such unpleasant subjects."
3. A doctor comforts his patient: "You have a very serious disease. Of 10 persons who catch it, only one survives. It is lucky you came to me, for I have recently had nine patients with this disease and they all died of it."
4. Dialogue in a French film:
   "Sir, I would like to ask for your daughter's hand."
   "Why not? You have already had the rest."
5. A marquis of the court of Louis XV unexpectedly returned from a journey and, on entering his wife's boudoir, found her in the arms of a bishop. After a moment's hesitation, the marquis walked calmly to the window, leaned out, and began going through the motions of blessing the people in the street.
   "What are you doing?" cried the anguished wife.
   "Monseigneur is performing my functions, so I am performing his."

Is there a common pattern underlying these five stories? Starting with the last, a little reflection reveals that the marquis's behaviour is both unexpected and perfectly logical—but of a logic not usually applied to this type of situation. It is the logic of the division of labour, governed by rules as old as human civilization. But his reactions would have been expected to be governed by a different set of rules—the code of sexual morality. It is the sudden clash between these two mutually exclusive codes of rules—or associative contexts—that produces the comic effect. It compels the listener to perceive the situation in two self-consistent but incompatible frames of reference at the same time; his mind has to operate simultaneously on two different wavelengths. While this unusual condition lasts, the event is not only, as is normally the case, associated with a single frame of reference, but "disociated" with two. The word disociation was coined by the present writer to make a distinction between the routines of disciplined thinking within a single universe of discourse—on a single plane, as it were—and the creative types of mental activity that always operate on more than one plane. In humour, both the *creation* of a subtle joke and the *recreative* act of perceiving the joke involve the delightful mental jolt of a sudden leap from one plane or associative context to another.

Bisociation

Turning to the other examples, in the French film dialogue, the daughter's "hand" is perceived first in a metaphorical frame of reference, then suddenly in a literal, bodily context. The doctor thinks in terms of abstract, statistical probabilities, the rules of which are inapplicable to individual cases; and there is an added twist because, in contrast to what common sense suggests, the patient's odds of survival are unaffected by whatever happened before; they are still one against 10. This is one of the profound paradoxes of the theory of probability, and the joke in fact implies a riddle; it pinpoints an absurdity that tends to be taken for granted. As for the lady who looks upon death as "eternal bliss" and at the same time "an unpleasant subject," she epitomizes the common human predicament of living in the divided house of faith and reason. Here again the simple joke carries unconscious overtones and undertones, audible to the inner ear alone.

The masochist who punishes himself by depriving himself of his daily punishment is governed by rules that are a *reversal* of those of normal logic. (A pattern can be constructed in which *both* frames of reference are reversed: "A sadist is a person who is kind to a masochist.") But there

is again an added twist. The joker does not really believe that the masochist takes his hot shower as a punishment; he only pretends to believe it. *Irony* is the satirist's most effective weapon; it pretends to adopt the opponent's ways of reasoning in order to expose their implicit absurdity or viciousness.

The common pattern underlying these stories is *the perceiving of a situation in two self-consistent but mutually incompatible frames of reference or associative contexts.* This formula can be shown to have a general validity for all forms of humour and wit, some of which will be discussed below. But it covers only one aspect of humour—its *intellectual structure.* Another fundamental aspect must be examined—the *emotional dynamics* that breathe life into that structure and make a person laugh, giggle, or smile.

LAUGHTER AND EMOTION

When a comedian tells a story, he deliberately sets out to create a certain tension in his listeners, which mounts as the narrative progresses. But it never reaches its expected climax. The punch line, or point, acts as a verbal guillotine that cuts across the logical development of the story; it debunks the audience's dramatic expectations. The tension that was felt becomes suddenly redundant and is exploded in laughter. To put it differently, laughter disposes of emotive excitations that have become pointless and must somehow be worked off along physiological channels of least resistance; and the function of the "luxury reflex" is to provide these channels.

A glance at the caricatures of the 18th-century English artists William Hogarth or Thomas Rowlandson, showing the brutal merriment of people in a tavern, makes one realize at once that they are working off their surplus of adrenalin by contracting their face muscles into grimaces, slapping their thighs, and breathing in puffs through the half-closed glottis. Their flushed faces reveal that the emotions disposed of through these safety valves are brutality, envy, sexual gloating. In cartoons by the 20th-century American James Thurber, however, coarse laughter yields to an amused and rarefied smile: the flow of adrenalin has been distilled and crystallized into a grain of Attic salt—a sophisticated joke. The word witticism is derived from "wit" in its original sense of intelligence and acumen (as is *Witz* in German). The domains of humour and of ingenuity are continuous, without a sharp boundary: the jester is brother to the sage. Across the spectrum of humour, from its coarse to its subtle forms, from practical joke to brainteaser, from jibe to irony, from anecdote to epigram, the emotional climate shows a gradual transformation. The emotion discharged in coarse laughter is aggression robbed of its purpose. The jokes small children enjoy are mostly scatological; adolescents of all ages gloat on vicarious sex. The sick joke trades on repressed sadism, satire on righteous indignation. There is a bewildering variety of moods involved in different forms of humour, including mixed or contradictory feelings; but whatever the mixture, it must contain a basic ingredient that is indispensable: an impulse, however faint, of aggression or apprehension. It may appear in the guise of malice, contempt, the veiled cruelty of condescension, or merely an absence of sympathy with the victim of the joke—a momentary anesthesia of the heart, as the French philosopher Henri Bergson put it.

The aggressive element

In the subtler types of humour, the aggressive tendency may be so faint that only careful analysis will detect it, like the presence of salt in a well-prepared dish—which, however, would be tasteless without it. Replace aggression by sympathy and the same situation—a drunk falling on his face, for example—will be no longer comic but pathetic and will evoke not laughter but pity. It is the aggressive element, the detached malice of the comic impersonator, that turns pathos into bathos, tragedy into travesty. Malice may be combined with affection in friendly teasing; and the aggressive component in civilized humour may be sublimated or no longer conscious. But in jokes that appeal to children and primitive people, cruelty and boastful self-assertiveness are much in evidence. In 1961 a survey carried out among American children aged eight to 15 made the researchers conclude that the mortification, discomfort, or hoaxing of others readily caused laughter, but witty or funny remarks often passed unnoticed.

Theories of laughter

Similar considerations apply to the historically earlier forms and theories of the comic. In Aristotle's view, laughter was intimately related to ugliness and debasement. Cicero held that the province of the ridiculous lay in a certain baseness and deformity. Descartes believed that laughter was a manifestation of joy mixed with surprise or hatred or both. In Francis Bacon's list of what causes laughter, the first place is again given to deformity. One of the most frequently quoted utterances on the subject is this definition in Thomas Hobbes's *Leviathan* (1651):

> The passion of laughter is nothing else but sudden glory arising from a sudden conception of some eminency in ourselves by comparison with the infirmity of others, or with our own formerly.

In the 19th century, Alexander Bain, an early experimental psychologist, thought along the same lines:

> Not in physical effects alone, but in everything where a man can achieve a stroke of superiority, in surpassing or discomforting a rival, is the disposition of laughter apparent.

In Bergson's view, laughter is the corrective punishment inflicted by society upon the unsocial individual: "In laughter we always find an unavowed intention to humiliate and consequently to correct our neighbour." Sir Max Beerbohm, the 20th-century English wit, found "two elements in the public's humour: delight in suffering, contempt for the unfamiliar." The American psychologist William McDougall believed that "laughter has been evolved in the human race as an antidote to sympathy, a protective reaction shielding us from the depressive influence of the shortcomings of our fellow men."

However much the opinions of the theorists differ, on this one point nearly all of them agree: that the emotions discharged in laughter always contain an element of aggressiveness. It must be borne in mind, however, that aggression and apprehension are twin phenomena, so much so that psychologists are used to talking of "aggressive–defensive impulses." Accordingly, one of the typical situations in which laughter occurs is the moment of sudden cessation of fear caused by some imaginary danger. Rarely is the nature of laughter as an overflow of redundant tensions more strikingly manifested than in the sudden change of expression on a small child's face from anxious apprehension to the happy laughter of relief. This seems to be unrelated to humour; yet a closer look reveals in it the same logical structure as in the joke: the wildly barking little dog was first perceived by the child in a context of danger, then discovered to be a harmless pup; the tension has suddenly become redundant and is spilled.

Laughter as the release of tension

Immanuel Kant realized that what causes laughter is "the sudden transformation of a tense expectation into nothing." Herbert Spencer, the 19th-century English philosopher, took up the idea and attempted to formulate it in physiological terms: "Emotions and sensations tend to generate bodily movements. . . . When consciousness is unawares transferred from great things to small," the "liberated nerve force" will expend itself along channels of least resistance—the bodily movements of laughter. Freud incorporated Spencer's theory of humour into his own, with special emphasis on the release of repressed emotions in laughing; he also attempted to explain why the excess energy should be worked off in that particular way:

> According to the best of my knowledge, the grimaces and contortions of the corners of the mouth that characterise laughter appear first in the satisfied and over-satiated nursling when he drowsily quits the breast. . . . They are physical expressions of the determination to take no more nourishment, an "enough" so to speak, or rather a "more than enough" . . . This primal sense of pleasurable saturation may have provided the link between the smile—that basic phenomenon underlying laughter—and its subsequent connection with other pleasurable processes of de-tension.

In other words, the muscle contractions of the smile, as the earliest expressions of relief from tension, would thereafter serve as channels of least resistance. Similarly, the explosive exhalations of laughter seem designed to "puff away" surplus tension in a kind of respiratory gymnastics, and agitated gestures obviously serve the same function.

It may be objected that such massive reactions often seem quite out of proportion to the slight stimulations that provoke them. But it must be borne in mind that laughter is a phenomenon of the trigger-releaser type, where a sudden turn of the tap may release vast amounts of stored emotions, derived from various, often unconscious, sources: repressed sadism, sexual tumescence, unavowed fear, even boredom. The explosive laughter of a class of schoolboys at some trivial incident is a measure of their pent-up resentment during a boring lecture. Another factor that may amplify the reaction out of all proportion to the comic stimulus is the social infectiousness that laughter shares with other emotive manifestations of group behaviour.

Patterns of association

Laughter or smiling may also be caused by stimulations that are not in themselves comic but signs or symbols deputizing for well-established comic patterns—such as Charlie Chaplin's oversized shoes or Groucho Marx's cigar—or catchphrases, or allusions to family jokes. To discover why people laugh requires, on some occasions, tracing back a long, involved thread of associations to its source. This task is further complicated by the fact that the effect of such comic symbols—in a cartoon or on the stage—appears to be instantaneous, without allowing time for the accumulation and subsequent discharge of "expectations" and "emotive tensions." But here memory comes into play, having already accumulated the required emotions in past experiences, acting as a storage battery whose charge can be sparked off at any time: the smile that greets Falstaff's appearance on the scene is derived from a mixture of memories and expectations. Besides, even if a reaction to a cartoon appears to be instantaneous, there is always a process in time until the reader "sees the joke"; the cartoon has to tell a story even if it is telescoped into a few seconds. All of this shows that to analyze humour is a task as delicate as analyzing the composition of a perfume with its multiple ingredients, some of which are never consciously perceived while others, when sniffed in isolation, would make one wince.

In this article there has been a discussion first of the logical structure of humour and then of its emotional dynamics. Putting the two together, the result may be summarized as follows: the "bisociation" of a situation or idea with two mutually incompatible contexts in a person's mind and the resulting abrupt transfer of his train of thought from one context to another put a sudden end to his "tense expectations"; the accumulated emotion, deprived of its object, is left hanging in the air and is discharged in laughter. Upon hearing that the marquis in the story told earlier walks to the window and starts blessing the people in the street, the intellect turns a somersault and enters with gusto into the new game. The malicious and erotic feelings aroused by the start of the story, however, cannot be fitted into the new context; deserted by the nimble intellect, these feelings gush out in laughter like air from a punctured tire.

Separation of thought and emotion

To put it differently: people laugh because their emotions have a greater inertia and persistence than their thoughts. Affects are incapable of keeping step with reasoning; unlike reasoning, they cannot "change direction" at a moment's notice. To the physiologist, this is self-evident since emotions operate through the genetically old, massive sympathetic nervous system and its allied hormones, acting on the whole body, while the processes of conceptual thinking are confined to the neocortex at the roof of the brain. Common experience provides daily confirmation of this dichotomy. People are literally "poisoned" by their adrenal humours; it takes time to talk a person out of a mood; fear and anger show physical aftereffects long after their causes have been removed. If man were able to change his moods as quickly as his thoughts, he would be an acrobat of emotion; but since he is not, his thoughts and emotions frequently become dissociated. It is emotion deserted by thought that is discharged in laughter. For emotion, owing to its greater mass momentum, is, as has been shown, unable to follow the sudden switch of ideas to a different type of logic; it tends to persist in a straight line. Aldous Huxley once wrote:

> We carry around with us a glandular system which was admirably well adapted to life in the Paleolithic times but is not very well adapted to life now. Thus we tend to produce more adrenalin than is good for us, and we either suppress ourselves and turn destructive energies inwards or else we do not suppress ourselves and we start hitting people. (From *Man and Civilization: Control of the Mind,* ed. Seymour M. Farber and Roger H.L. Wilson. Copyright 1961. Used with permission of McGraw-Hill Book Company.)

A third alternative is to laugh at people. There are other outlets for tame aggression, such as competitive sports or literary criticism; but they are acquired skills, whereas laughter is a gift of nature, included in man's native equipment. The glands that control his emotions reflect conditions at a stage of evolution when the struggle for existence was more deadly than at present—and when the reaction to any strange sight or sound consisted in jumping, bristling, fighting, or running. As security and comfort increased in the species, new outlets were needed for emotions that could no longer be worked off through their original channels, and laughter is obviously one of them. But it could only emerge when reasoning had gained a degree of independence from the urges of emotion. Below the human level, thinking and feeling appear to form an indivisible unity. Not before thinking became gradually detached from feeling could man perceive his own emotion as redundant and make the smiling admission, "I have been fooled."

VERBAL HUMOUR

The foregoing discussion was intended to provide the tools for dissecting and analyzing any specimen of humour. The procedure is to determine the nature of the two (or more) frames of reference whose collision gives rise to the comic effect—to discover the type of logic or "rules of the game" that govern each. In the more sophisticated type of joke, the logic is implied and hidden, and the moment it is stated in explicit form, the joke is dead. Unavoidably, the section that follows will be strewn with cadavers.

Puns

Max Eastman, in *The Enjoyment of Laughter* (1936), remarked of a laboured pun by Ogden Nash: "It is not a pun but a punitive expedition." That applies to most puns, including Milton's famous lines about the Prophet Elijah's ravens, which were "though ravenous taught to abstain from what they brought," or the character mentioned by Freud, who calls the Christmas season the "alcoholidays." Most puns strike one as atrocious, perhaps because they represent the most primitive form of humour; two disparate strings of thought tied together by an acoustic knot. But the very primitiveness of such association based on pure sound ("hol") may account for the pun's immense popularity with children and its prevalence in certain types of mental disorder ("punning mania").

From the play on sounds—puns and Spoonerisms—an ascending series leads to the play on words and so to the play on ideas. When Groucho Marx says of a safari in Africa, "We shot two bucks, but that was all the money we had," the joke hinges on the two meanings of the word buck. It would be less funny without the reference to Groucho, which evokes a visual image instantly arousing high expectations. The story about the marquis above may be considered of a superior type of humour because it plays not on mere words but on ideas.

It would be quite easy—and equally boring—to draw up a list in which jokes and witticisms are classified according to the nature of the frames of reference whose collision creates the comic effect. A few have already been mentioned: metaphorical versus literal meaning (the daughter's "hand"); professional versus common sense logic (the doctor); incompatible codes of behaviour (the marquis); confrontations of the trivial and the exalted ("eternal bliss"); trains of reasoning travelling, happily joined together, in opposite directions (the sadist who is kind to the masochist). The list could be extended indefinitely; in fact *any* two frames of reference can be made to yield a comic effect of sorts by hooking them together and infusing a drop of malice into the concoction. The frames may even be defined by such abstract concepts as "time" and "weather": the absent-minded professor who tries to read the temperature from his watch or to tell the time from the thermometer is comic in the same way as a game of

table tennis played with a soccer ball or a game of rugby played with a table tennis ball. The variations are infinite, the formula remains the same.

Jokes and anecdotes have a single point of culmination. The literary forms of *sustained humour,* such as the picaresque novel, do not rely on a single effect but on a series of minor climaxes. The narrative moves along the line of intersection of contrasted planes, such as the fantasy world of Don Quixote and the cunning horse sense of Sancho Panza, or is made to oscillate between them. As a result, tension is continuously generated and discharged in mild amusement.

Comic verse

Comic verse thrives on the melodious union of incongruities, such as the "cabbages and kings" in Lewis Carroll's "The Walrus and the Carpenter," and particularly on the contrast between lofty form and flat-footed content. Certain metric forms associated with heroic poetry, such as the hexameter or Alexandrine, arouse expectations of pathos, of the exalted; to pour into these epic molds some homely, trivial content—"beautiful soup, so rich and green/ waiting in a hot tureen"—is an almost infallible comic device. The rolling rhythms of the first lines of a limerick that carry, instead of a mythical hero such as Hector or Achilles, a young lady from Ohio for a ride make her ridiculous even before the expected calamities befall her. Instead of a heroic mold, a soft lyrical one may also pay off:

> . . . And what could be moister
> Than tears of an oyster?

Another type of incongruity between form and content yields the bogus proverb: "The rule is: jam tomorrow and jam yesterday—but never jam today." Two contradictory statements have been telescoped into a line whose homely, admonitory sound conveys the impression of a popular adage. In a similar way, nonsense verse achieves its effect by pretending to make sense, by forcing the reader to project meaning into the phonetic pattern of the jabberwocky, as one interprets the ink blots in a Rorschach test.

Satire and allegory

The *satire* is a verbal caricature that shows a deliberately distorted image of a person, institution, or society. The traditional method of the caricaturist is to exaggerate those features he considers to be characteristic of his victim's personality and to simplify by leaving out everything that is not relevant for his purpose. The satirist uses the same technique, and the features of society he selects for magnification are, of course, those of which he disapproves. The result is a juxtaposition, in the reader's mind, of his habitual image of the world in which he moves and its absurd reflection in the satirist's distorting mirror. He is made to recognize familiar features in the absurd and absurdity in the familiar. Without this double vision the satire would be humourless. If the human Yahoos were really such evil-smelling monsters as Gulliver's Houyhnhnm hosts claim, then Jonathan Swift's *Gulliver's Travels* (1726) would not be a satire but the statement of a deplorable truth. Straight invective is not satire; satire must deliberately overshoot its mark.

A similar effect is achieved if, instead of exaggerating the objectionable features, the satirist projects them by means of the *allegory* onto a different background, such as an animal society. A succession of writers, from the ancient Greek dramatist Aristophanes through Swift to such 20th-century satirists as Anatole France and George Orwell, have used this technique to focus attention on deformities of society that, blunted by habit, are taken for granted.

SITUATIONAL HUMOUR

The coarsest type of humour is the practical joke: pulling away the chair from under the dignitary's lowered bottom. The victim is perceived first as a person of consequence, then suddenly as an inert body subject to the laws of physics: authority is debunked by gravity, mind by matter; man is degraded to a mechanism. Goose-stepping soldiers acting like automatons, the pedant behaving like a mechanical robot, the Sergeant Major attacked by diarrhea, or Hamlet getting the hiccups—all show man's lofty aspirations deflated by his all-too-solid flesh. A similar effect is produced by artifacts that masquerade as humans: Punch and Judy, jack-in-the-box, gadgets playing tricks on their masters as if with calculated malice.

In Henri Bergson's theory of laughter, this dualism of subtle mind and inert matter—he calls it "the mechanical encrusted on the living"—is made to serve as an explanation of *all* varieties of the comic. In the light of what has been said, however, it would seem to apply only to *one* type of comic situation among many others.

From the "bisociation" of man and machine, there is only a step to the man–animal hybrid. Walt Disney's creations behave as if they were human without losing their animal appearance. The caricaturist follows the reverse procedure by discovering horsey, mousy, or piggish features in the human face.

This leads to the comic devices of imitation, impersonation, and disguise. The impersonator is perceived as himself and somebody else at the same time. If the result is slightly degrading—but only in that case—the spectator will laugh. The comedian impersonating a public personality, two pairs of trousers serving as the legs of the pantomime horse, men disguised as women and women as men—in each case the paired patterns reduce each other to absurdity.

The most aggressive form of impersonation is the *parody,* designed to deflate hollow pretense, to destroy illusion, and to undermine pathos by harping on the weaknesses of the victim. Wigs falling off, speakers forgetting their lines, gestures remaining suspended in the air: the parodist's favourite points of attack are again situated on the line of intersection between the sublime and the trivial.

Playful behaviour in young animals and children is amusing because it is an unintentional parody of adult behaviour, which it imitates or anticipates. Young puppies are droll because their helplessness, affection, and puzzled expression make them appear more "human" than full-grown dogs; because their growls strike one as impersonations of adult behaviour—like a child in a bowler hat; because the puppy's waddling, uncertain gait makes it a choice victim of nature's practical jokes; because its bodily disproportions—the huge padded paws, Falstaffian belly, and wrinkled brow—give it the appearance of a caricature; and lastly because the observer feels so superior to a puppy. A fleeting smile can contain many logical ingredients and emotional spices.

Both Cicero and Francis Bacon regarded *deformity* as the most frequent cause of laughter. Renaissance princes collected dwarfs and hunchbacks for their merriment. It obviously requires a certain amount of imagination and empathy to recognize in a midget a fellow human, who, though different in appearance, thinks and feels much as oneself does. In children, this projective faculty is still rudimentary: they tend to mock people with a stammer or a limp and laugh at the foreigner with an odd pronunciation. Similar attitudes are shown by tribal or parochial societies to any form of appearance or behaviour that deviates from their strict norms: the stranger is not really human; he only pretends to be "like us." The Greeks used the same word, barbarous, for the foreigner and the stutterer: the uncouth barking sounds the stranger uttered were considered a parody of human speech. Vestiges of this primitive attitude are still found in the curious fact that civilized people accept a foreign accent with tolerance, whereas imitation of a foreign accent strikes them as comic. The imitator's mispronunciations are recognized as mere pretense; this knowledge makes sympathy unnecessary and enables the audience to be childishly cruel with a clean conscience.

Other sources of innocent laughter are situations in which *the part and the whole* change roles, and attention becomes focussed on a detail torn out of the functional context on which its meaning depended. When the phonograph needle gets stuck, the soprano's voice keeps repeating the same word on the same quaver, which suddenly assumes a grotesquely independent life. The same happens when faulty orthography displaces attention from meaning to spelling, or whenever consciousness is directed at functions that otherwise are performed automatically. The latter situation is well illustrated by the story of the centipede who, when asked in which order he moved his

hundred legs, became paralyzed and could walk no more. The self-conscious, awkward youth, who does not know what to do with his hands, is a victim of the paradox of the centipede.

Comedies have been classified according to their reliance on situations, manners, or characters. The logic of the last two needs no further discussion; in the first, comic effects are contrived by making a situation participate simultaneously in two independent chains of events with different associative contexts, which intersect through coincidence, mistaken identity, or confusions of time and occasion.

Tickling

Why tickling should produce laughter remained an enigma in all earlier theories of the comic. As Darwin was the first to point out, the innate response to tickling is squirming and straining to withdraw the tickled part—a defense reaction designed to escape attacks on vulnerable areas such as the soles of the feet, armpits, belly, and flank. If a fly settles on the belly of a horse, it causes a ripple of muscle contractions across the skin—the equivalent of squirming in the tickled child. But the horse does not laugh when tickled, and the child not always. The child will laugh only—and this is the crux of the matter—when it perceives tickling as a *mock attack,* a caress in mildly aggressive disguise. For the same reason, people laugh only when tickled by others, not when they tickle themselves.

Experiments at Yale University on babies under one year revealed the not very surprising fact that they laughed 15 times more often when tickled by their mothers than by strangers; and when tickled by strangers, they mostly cried. For the mock attack must be recognized as being only pretense, and with strangers one cannot be sure. Even with its own mother, there is an ever-so-slight feeling of uncertainty and apprehension, the expression of which will alternate with laughter in the baby's behaviour. It is precisely this element of tension between the tickles that is relieved in the laughter accompanying the squirm. The rule of the game is "let me be just a little frightened so that I can enjoy the relief."

Thus the tickler is impersonating an aggressor but is simultaneously known not to be one. This is probably the first situation in life that makes the infant live on two planes at once, a delectable foretaste of being tickled by the horror comic.

Humour in the visual arts and music

Humour in the visual arts reflects the same logical structures as discussed before. Its most primitive form is the distorting mirror at the fun fair, which reflects the human frame elongated into a column or compressed into the shape of a toad. It plays a practical joke on the victim, who sees the image in the mirror both as his familiar self and as a lump of plasticine that can be stretched and squeezed into any absurd form. The mirror distorts mechanically while the caricaturist does so selectively, employing the same method as the satirist—exaggerating characteristic features and simplifying the rest. Like the satirist, the caricaturist reveals the absurd in the familiar; and, like the satirist, he must overshoot his mark. His malice is rendered harmless by the knowledge that the monstrous potbellies and bowlegs he draws are not real; real deformities are not comic but arouse pity.

The artist, painting a stylized portrait, also uses the technique of selection, exaggeration, and simplification; but his attitude toward the model is usually dominated by positive empathy instead of negative malice, and the features he selects for emphasis differ accordingly. In some character studies by Leonardo da Vinci, Hogarth, or Honoré Daumier, the passions reflected are so violent, the grimaces so ferocious, that it is impossible to tell whether the works were meant as portraits or caricatures. If one feels that such distortions of the human face are not really possible, that Daumier merely *pretended* that they exist, then one is absolved from horror and pity and can laugh at his grotesques. But if one feels that this is indeed what Daumier saw in those dehumanized faces, then they are not comic but tragic.

Humour in *music* is a subject to be approached with diffidence because the language of music ultimately eludes translation into verbal concepts. All one can do is to point out some analogies: a "rude" noise, such as the blast of a trumpet inserted into a passage where it does not belong, has the effect of a practical joke; a singer or an instrument out of tune produces a similar reaction; the imitation of animal sounds, vocally or instrumentally, exploits the technique of impersonation; a nocturne by Chopin transposed into hot jazz or a simple street song performed with Wagnerian pathos is a marriage of incompatibles. These are primitive devices corresponding to the lowest levels of humour; more sophisticated are the techniques employed by Maurice Ravel in *La Valse,* a parody of the sentimental Viennese waltz, or by Zoltán Kodály in the mock-heroics of his Hungarian folk opera, *Háry János.* But in comic operas it is almost impossible to sort out how much of the comic effect is derived from the book and how much from the music; and the highest forms of musical humour, the unexpected delights of a lighthearted scherzo by Mozart, defy verbal analysis, unless it is so specialized and technical as to defeat its purpose. Although a "witty" musical passage that springs a surprise on the audience and cheats it of its expectations certainly has the emotion-relieving effect that tends to produce laughter, a concert audience may occasionally smile but will hardly ever laugh: the emotions evoked by musical humour are of a subtler kind than those of the verbal and visual variety.

STYLES AND TECHNIQUES IN HUMOUR

The criteria that determine whether a humorous offering will be judged good, bad, or indifferent are partly a matter of period taste and personal preference and partly dependent on the style and technique of the humorist. It would seem that these criteria can be summed up under three main headings: originality, emphasis, and economy.

Originality, emphasis, and economy

The merits of originality are self-evident; it provides the essential element of surprise, which cuts across our expectations. But true originality is not very often met either in humour or in other forms of art. One common substitute for it is to increase the tension of the audience by various techniques of suggestive emphasis. The clown's domain is the rich, coarse type of humour: he piles it on; he appeals to sadistic, sexual, scatological impulses. One of his favourite tricks is repetition of the same situation, the same key phrase. This diminishes the effect of surprise, but it has a tension-accumulating effect: emotion is easily drawn into the familiar channel—more and more liquid is being pumped into the punctured pipeline.

Emphasis on local colour and ethnic peculiarities, such as Scottish or Cockney stories, for example, is a further means to channel emotion into familiar tracks. The Scotsman or Cockney must, of course, be a caricature if the comic purpose is to be achieved. In other words, exaggeration and simplification once more appear as indispensable tools to provide emphasis.

In the higher forms of humour, however, emphasis tends to yield to the opposite kind of virtue—economy. Economy, in humour and art, does not mean mechanical brevity but implicit hints instead of explicit statements—the oblique allusion in lieu of the frontal attack. Old-fashioned cartoons, such as those featuring the British lion and the Russian bear, hammered their message in; the modern cartoon usually poses a riddle that the reader must solve by an imaginative effort in order to see the joke.

In humour, as in other forms of art, emphasis and economy are complementary techniques. The first forces the offering down the consumer's throat; the second tantalizes to whet his appetite.

RELATIONS TO ART AND SCIENCE

Earlier theories of humour, including even those of Bergson and Freud, treated it as an isolated phenomenon, without attempting to throw light on the intimate connections between the comic and the tragic, between laughter and crying, between artistic inspiration, comic inventiveness, and scientific discovery. Yet these three domains of creative activity form a continuum with no sharp boundaries between wit and ingenuity, nor between discovery and art.

It has been said that scientific discovery consists in seeing an analogy where nobody has seen one before. When, in the Song of Solomon, Solomon compared the Shulamite's neck to a tower of ivory, he saw an analogy that nobody had seen before; when William Harvey compared

the heart of a fish to a mechanical pump, he did the same; and when the caricaturist draws a nose like a cucumber, he again does just that. In fact, all the logical patterns discussed above, which constitute a "grammar" of humour, can also enter the service of art or discovery, as the case may be. The pun has structural equivalents in the rhyme and in word games, which range from crossword puzzles to the deciphering of the Rosetta Stone, the key to Egyptian hieroglyphic. The confrontation between diverse codes of behaviour may yield comedy, tragedy, or new psychological insights. The dualism of mind and inert matter is exploited by the practical joker but also provides one of the eternal themes of literature: man as a marionette on strings, manipulated by gods or chromosomes. The man–beast dichotomy is reflected by Walt Disney's cartoon character Donald Duck but also in Franz Kafka's macabre tale *The Metamorphosis* (1915) and in the psychologist's experiments with rats. The caricature corresponds not only to the artist's character portrait but also to the scientist's diagrams and charts, which emphasize the relevant features and leave out the rest.

Synthesis, juxtaposition, and collision

Contemporary psychology regards the conscious and unconscious processes underlying creativity in all domains as an essentially combinative activity—the bringing together of previously separate areas of knowledge and experience. The scientist's purpose is to achieve *synthesis;* the artist aims at a *juxtaposition* of the familiar and the eternal; the humorist's game is to contrive a *collision.* And as their motivations differ, so do the emotional responses evoked by each type of creativity: discovery satisfies the exploratory drive; art induces emotional catharsis; humour arouses malice and provides a harmless outlet for it. Laughter has been described as the "Haha reaction"; the discoverer's Eureka cry as the "Aha! reaction"; and the delight of the aesthetic experience as the "Ah . . . reaction." But the transitions from one to the other are continuous: witticism blends into epigram, caricature into portrait; and whether one considers architecture, medicine, chess, or cookery, there is no clear frontier where the realm of science ends and that of art begins: the creative person is a citizen of both. Comedy and tragedy, laughter and weeping, mark the extremes of a continuous spectrum, and a comparison of the physiology of laughter and weeping yields further clues to this challenging problem. Such considerations, however, lie beyond the terms of reference of the present article.

THE HUMANIZATION OF HUMOUR

The San (Bushmen) of the Kalahari desert of South West Africa/Namibia are among the oldest and most primitive inhabitants of the Earth. An anthropologist who made an exhaustive study of them provided a rare glimpse of prehistoric humour:

> On the way home we saw and shot a springbok, as there was no meat left in camp. The bullet hit the springbok in the stomach and partly eviscerated him, causing him to jump and kick before he finally died. The Bushmen thought that this was terribly funny and they laughed, slapping their thighs and kicking their heels to imitate the springbok, showing no pity at all, but then they regard animals with great detachment.

But the San remained "in good spirits, pleased with the amusement the springbok had given them." (From Elizabeth Marshall Thomas, *The Harmless People;* Alfred A. Knopf, New York, 1959.)

Obviously the San, like most primitive people, do not regard animals as sentient beings; the springbok's kicking in his agony appears to them funny because in their view the animal *pretends* to suffer pain like a human being, though it is incapable of such feelings. The ancient Greeks' attitude toward the stammering barbarian was similarly inspired by the conviction that he is not really human but only pretends to be. The ancient Hebrews' sense of humour seems to have been no less harsh: it has been pointed out that in the Old Testament there are 29 references to laughter, out of which 13 instances are linked with scorn, derision, mocking, and contempt and only two are born of joy.

As laughter emerged from antiquity, it was so aggressive that it has been likened to a dagger. It was in ancient Greece that the dagger was transformed into a quill, dripping with poison at first, then diluted and infused with delightfully lyrical and fanciful ingredients. The 5th century BC saw the first rise of humour into art, starting with parodies of Olympian heroics and soon reaching a peak, in some respects unsurpassed to this day, in the comedies of Aristophanes. From here onward, the evolution of humour in the Western world merges with the history of literature and art.

If the overall trend was toward the humanization of humour from primitive to sophisticated forms, there also have been ups and downs reflecting changes in political and cultural climate. George Orwell's satire of the 20th century, for example, is much more savage than that of Jonathan Swift in 18th-century England or of Voltaire in 18th-century France. If the Dark Ages produced works of humorous art, little of it has survived. And under the tyrannies of Hitler in Germany and of Stalin in the Soviet Union, humour was driven underground. Dictators fear laughter more than bombs.

Non-Western styles. About non-Western varieties of humour, the Westerner is tempted to repeat the middle-aged British matron's remark on watching Cleopatra rave and die on the stage: "How different, how very different from the home life of our dear Queen." Humour thrives only in its native climate, embedded in its native logic; when one does not know what to expect, one cannot be cheated of one's expectations. Hindu humour, for instance, as exemplified by the savage pranks played on humans by the monkey-god Hanuman, strikes the Westerner as particularly cruel, perhaps because the Hindu's approach to mythology is fundamentally alien to the Western mind.

Japanese humour

The humour of the Japanese, on the other hand, is, from the Western point of view, astonishingly mild and poetical, like weak, mint-flavoured tea:

> The boss of the monkeys ordered his thousands of henchmen to get the moon reflected in the water. They all tried various means but failed and were much troubled. One of the monkeys at last got the moon in the water and respectfully offered it to the boss, saying "This is what you asked for." The boss was delighted and praised him, saying, "What an exploit! You have distinguished yourself!" The monkey then asked, "By the way, master, what are you going to do with this?" "Well, yes . . . I didn't think of that." (From *Karukuchi Ukibyotan,* 1751; in R.H. Blyth, *Japanese Humour,* 1957.)

The following dates from about a century later:

> There was once a man who was always bewailing his lack of money to buy saké (rice wine) with. His wife, feeling sorry for him, dutifully cut off some of her hair and sold it to the hairdresser's for twenty-four mon, and bought her husband some saké. "Where on earth did you get this from?" "I sold my hair and bought it." "You did such a thing for me?" The wretched man shed tears, and fondling his wife's remaining hair said, "Yes, and there's another good half-bottle of saké here!" (From *Chanoko-mochi,* 1856; in Blyth.)

The combination of maudlin tears and brazen selfishness, and the crazy logic of equating the wife's coiffure with a liquid measure of saké, show the familiar Western pattern of the clash of incompatibles, even though transplanted into another culture.

Humour in the contemporary world. Humour today seems to be dominated by two main factors: the influence of the mass media and the crisis of values affecting a culture in rapid and violent transition. The former tends toward the commercialized manufacture of laughter by popular comedians and gags produced by conveyor-belt methods; the latter toward a sophisticated form of black humour larded with sick jokes, sadism, and sex.

Fashions, however, always run their course; perhaps the next one will delight in variations on the theme of the monkey boss who, having gained possession of the moon, does not know what to do with it. The only certainty regarding the humour of the future is contained in Dr. Samuel Johnson's dictum: "Sir, men have been wise in many different modes, but they have always laughed in the same way."

BIBLIOGRAPHY. GUILLAUME DUCHENNE (DE BOULOGNE), *Le Mécanisme de la physionomie humaine* (1862); and JULES-MARIE RAULIN, *Le Rire et les exhilarantes: étude anatomique, psycho-physiologique et pathologique* (1900), contain valuable

source material on the study of the physiology of laughter, including experiments by the neurologist Jean-Martin Charcot and the Nobel laureate Charles Richet. HERBERT SPENCER, in his "The Physiology of Laughter," *Essays on Education and Kindred Subjects* (1910, reprinted 1977), outlined the tension-relieving function of humour, on which FREUD elaborated in his *Wit and Its Relations to the Unconscious* (1916; originally published in German, 1905), with special emphasis on infantile and repressed elements. HENRI BERGSON, *Laughter* (1911, reprinted 1937; originally published in French, 1900), is a classic work attempting to derive all types of humour from the contrast between mind and matter. MAX F. EASTMAN, *Enjoyment of Laughter* (1936), is unique in that he denies the malicious element in laughter. REGINALD H. BLYTH, *Japanese Humor,* 2nd ed. (1961), throws a delightful sidelight on a different culture. DAVID H. MONRO, *Argument of Laughter* (1951, reprinted 1963), contains a valuable summary of earlier theories. ARTHUR KOESTLER, *Insight and Outlook* (1949, reprinted 1965), and *The Act of Creation* (1964, reprinted 1976), attempt to present a synthetic theory of humour and its interrelations with art and discovery.

(A.Ko.)

Hungarian Literature

No written evidence remains of the earliest Hungarian literature, but through Hungarian folktales and folk songs elements have survived that can be traced back to pagan times. Also extant, although only in Latin and dating from between the 11th and 14th centuries, are shortened versions of some Hungarian legends relating the origins of the Hungarian people and episodes from the conquest of Hungary and from the Hungarian campaigns of the 10th century.

This article is divided into the following sections:

Earliest writings in Hungarian 689
Renaissance and Reformation 689
The 17th century 689
  Effects of the Counter-Reformation
  Period of decline
The period of the Enlightenment 690
The 19th century 691
  Romanticism
  Writers of the late 19th century
The 20th century 691
  Early years
  The interwar period
  Writing after 1945
Bibliography 692

EARLIEST WRITINGS IN HUNGARIAN

The earliest known written traces of the Hungarian language are mostly proper names embedded in the Latin text of legal or ecclesiastical documents. The first continuous example of the Hungarian language is the *Halotti beszéd,* a short funeral oration written in about 1200, moving in its simplicity. Many translations from Latin were made in the 13th and 14th centuries, but the only one that has survived, and also the oldest extant poem written in Hungarian, is a free version of a poem by Godefroy de Breteuil. It is known as *Ómagyar Mária-siralom* (*c.* 1300; "Old Hungarian Lament of the Virgin Mary"). The 14th century also produced translations of the legends of St. Margaret and St. Francis of Assisi. The *Jókai codex,* which contains the St. Francis legend, was written in about 1440 and is the oldest extant Hungarian codex.

The 15th century saw the first translations from the Bible. The preachers Thomas and Valentine, followers of the Bohemian religious reformer Jan Hus, were responsible for this work, of which the prophetic books, the Psalms, and the Gospels have survived. A great part of the vocabulary, created for the purpose, is still in use. A number of sermons by the Franciscan Pelbárt Temesvári, originally written in Latin, have come down in Hungarian translations. Among other translations are the first Hungarian drama, *A három körösztény leányról* (*c.* 1520; *On Three Christian Virgins*), translated from the Latin original of Hrosvitha; a translation of the legend of St. Catherine of Alexandria; and a translation of the Song of Solomon.

Hungarian literature was not entirely religious. The existence of a history of the Trojan War and of a Hungarian version of the Alexander romance can be inferred from their South Slavic translations. In the 14th century secular literature developed and literary forms were introduced from abroad.

RENAISSANCE AND REFORMATION

In 1367 the first Hungarian university was founded, at Pécs. About 100 years later King Matthias I Corvinus established the first Hungarian printing press. The King became known for his library and his patronage of foreign scholars; during his reign Latin literature in Hungary reached its peak in Janus Pannonius, who had been educated in Italy.

The 16th century brought changes. After the Battle of Mohács (1526) the Ottoman Turks occupied a large part of Hungary and the country was split into three. It is in the era of the Reformation that Hungarian national literature really began. Benedek Komjáti, Gábor Pesti, and János Sylvester, all of whom were disciples of the humanist Erasmus, translated parts of the Bible with philological accuracy. Pesti made a very readable translation of Aesop's fables and published a Latin–Hungarian dictionary. Sylvester published the first Hungarian grammar and, to show the adaptability of the vernacular to classical verse forms, wrote the first Hungarian poem in couplets. In 1541 he published a translation of the New Testament.

The Hungarian Reformation

The second half of the 16th century saw the beginnings of Hungarian drama. *Comoedia Balassi Menyhárt árultátásáról* (1569; "Comedy on the Treachery of Menyhárt Balassi"), a satire by an unknown author, was among the most interesting literary achievements of the Reformation. Péter Bornemisza, the first important Protestant writer in Hungary, gave an entrancing view of Hungarian life, teeming with fresh observations, vivid descriptions, and original comments. His volume *Ördögi Kisértetekről* (1578; "On the Temptations of the Devil") offered an interesting consideration of moral and sexual problems in the 16th century. A poem of farewell, written on leaving the country, was one of the gems of early Hungarian poetry. His *Tragoedia magyar nyelven* (1558; "Tragedy in Hungarian"), though based on Sophocles' *Electra,* is a skillful adaptation of the play in the spirit of humanism.

Perhaps the greatest single literary achievement of the Hungarian Reformation was a translation of the Bible by Gáspár Károlyi and others (1590). The translation played a role in the development of Hungarian similar to that of the Authorized Version in English.

Up to the 16th century religious literature seems to have fared better than secular literature, in part because secular literature was not written down. The late 16th-century minstrels were more learned than their predecessors and in many cases were driven to their profession by difficult economic conditions. Perhaps the most important was Sebestyén Tinódi, by temperament more historian than poet. He described the wars against the Turks with remarkable accuracy, but his verse was monotonous. Péter Ilosvai Selymes was the author of a romance, *Az híres nevezetes Toldi Miklósnak jeles cselekedetiről* (1574; "The Story of the Remarkable Nicholas Toldi's Extraordinary and Brave Deeds"), which achieved great popularity in Hungary and served as a basis for a masterpiece by János Arany in the 19th century. This romance was the one original piece in the flow of the mere entertainment literature characteristic of the 16th century, the principal genre of which was the *széphistória* ("beautiful story"), adapted from western European originals. Perhaps the best was the *História egy Árgirus nevű királyfiról* (*c.* 1575; "The Story of the Prince Argirus") by Albert Gergei, from an Italian original but interwoven with Hungarian folklore.

Bálint Balassi

A great poet emerged in Bálint Balassi (1554–94), who at first imitated Petrarch and various Neo-Latin poets but later displayed originality with a cycle of love poems of great beauty and emotional intensity. His songs of war, while reflecting the vicissitudes of fighting the Turk at the borders of the Christian world, celebrate nature and individual bravery in almost hymnlike tones. The poetry of his last years is imbued with a deep religious feeling; the imagery of the poems of his last creative period (the Coelie cycle of love songs as well as his religious verse) is coloured by Mannerism.

THE 17TH CENTURY

In the 17th century Hungary was still divided into three parts. The first, under Turkish rule, played no part in the

development of Hungarian literature. The second, under Habsburg rule, was open to Italian and German Roman Catholic influence; the third, Transylvania, was in close relationship with Dutch and English Protestant thought. The leading Protestant scholar and writer of the 17th century was János Apáczai Csere. His chief work was a Hungarian encyclopaedia in which he endeavoured to sum up the knowledge of his time. The work, published at Utrecht in 1653, marked a development in technical vocabulary.

Effects of the Counter-Reformation. By the end of the 16th century the Counter-Reformation was gaining momentum in western Hungary. A Jesuit cardinal, Péter Pázmány, a master of Hungarian prose, was outstanding as an orator and essayist. His writing was characterized by a vigorous and clear, though far from simple, style, use of popular expressions, and solid argument. His *Isteni igazságra vezérlő kalauz* (1613; "Guide to Divine Truth") was a refutation of non-Catholic religious doctrines and a masterpiece of Baroque prose.

Under the influence of the Jesuits, many Hungarian aristocrats returned to the Catholic faith and sent their sons to the Austrian Catholic universities and to Rome. The Italian Baroque, especially the influence of Tasso and Marino, is evident in the work of Miklós Zrínyi, a great Hungarian statesman and military commander. Most of his prose work was an exposition of political and strategic ideas. His greatest literary achievement was an epic, *Szigeti veszedelem* (1651; "The Peril of Sziget"), in 15 cantos, on the siege in 1566 of Szigetvár, which had been defended against the Turks by Zrínyi's great-grandfather. Though the influence of classical epics is clear, the work remains profoundly original and Hungarian. Another poet of this time, István Gyöngyösi, composed long narrative poems and also many epithalamia, or nuptial poems. He was inventive and handled rhyme with ease, and his work was read widely during the 17th and 18th centuries.

Period of decline. The period between 1700 and about 1770 was a time of decline and slow consolidation in Hungarian literature. Memoirs form what is best in the prose literature of the period. Among the most absorbing are the autobiography of the well-traveled Miklós Bethlen, a leading Transylvanian statesman, and the confessions and memoirs (written in Latin and French, respectively) of Ferenc Rákóczi II, exiled prince of Transylvania and leader of the anti-Habsburg insurrection of 1703–11. The *Törökországi levelek* ("Letters from Turkey"), written from 1717 to 1758 by Kelemen Mikes, a companion in exile of Rákóczi, were addressed to an imaginary aunt. In choosing the epistolary genre Mikes was inspired by French models, and his work stands out for its excellent style and wry humour. The *Metamorphosis Transylvaniae* of Péter Apor is a nostalgic reminiscence of Transylvania.

Memoirs and epistolary works

The poetry of this epoch has little to offer. The poems of László Amade were informed by a Rococo taste, both in form and in content; he mainly wrote poetry of gallantry and courtship. Another poet and translator, the versatile Ferenc Faludi, took his expressions from popular language and folk songs.

This period of literary decadence produced notable works only in the fields of history and history of literature. Some of them were written in Latin. Among historians, Mátyás Bél, György Pray, and István Katona are most important. The first historian of Hungarian literature, Dávid Czvittinger, composed the biographies of some 300 Hungarian writers. His work was continued and improved by Péter Bod, whose *Magyar Athénás* (1766; "Hungarian Athenaeum") deals with more than 500 Hungarian men of letters.

THE PERIOD OF THE ENLIGHTENMENT

The Hungarian Enlightenment was more receptive to French and English ideas than it was productive of original developments. The period between about 1772 and 1825, though immensely important in the development of the Hungarian spirit, produced few writers of the first rank.

With the publication in 1772 of the first literary work by György Bessenyei, a translation (from the French) of Alexander Pope's *Essay on Man,* the new era began. All of Bessenyei's works served a didactic purpose. His drama *Ágis tragédiája* (1772; "The Tragedy of Agis") was a somewhat creaking vehicle for his liberal ideas. His best work, *Tariménes utazása* (1802–04; "Tarimenes' Journey"), the first real novel in Hungarian, was a bitter attack on everything that was opposed to the Enlightenment. With destructive irony, Bessenyei, an officer of the Hungarian Guards, examined the shortcomings of contemporary society. His personal influence induced several of his fellow officers—for example, Sándor Báróczi and Ábrahám Barcsay—to try to convey the ideas of the Enlightenment in Hungarian to a Hungarian public.

Spurred on by new ideas, but basically traditionalists, József Gvadányi and András Dugonics produced amusing works that were both of some literary merit and popular. Gvadányi's best work, *Egy falusi nótáriusnak budai utazása* (1790; "The Journey to Buda of a Village Notary"), is a defense of national and traditional values against encroaching foreign ideas. The novel *Etelka* (1788), by Dugonics, a sentimental love story in a historical setting, was the first Hungarian best-seller. Both Gvadányi and Dugonics used the language of the common people, and this was perhaps their greatest merit. Ádám Pálóczi Horváth left a collection of 450 poems, a treasure-house of authentic folk songs.

The end of the 18th century was a period of experiments with poetic language. The pioneers of the use of Greek and Latin metres in Hungarian verse (to which they are eminently suited) were followed by Benedek Virág, who imbued with poetic inspiration verse forms that for his predecessors were merely formal exercises. It fell to Dániel Berzsenyi, who published a single volume of poetry, in 1813, to show what use a great poet could make of classical metre. His ode "A Magyarokhoz" ("To the Hungarians"), his "Fohász" ("Prayer"), and his elegy "A közelitő tél" ("On the Nearing Winter") express the transitoriness of power and of friendship.

The ideas of the Enlightenment were not universally welcomed in Hungary. Traditionalist elements looked with distrust on any imported ideas, and the government was increasingly suspicious of a spirit of intellectual freedom, which it believed had led to the French Revolution and, in Hungary, to the Jacobin conspiracy of Martinovics, crushed in 1794. Several writers went to prison for harbouring radical views. The most talented among them, János Batsányi, secured his place in the history of Hungarian literature by his poem "A Franciaországi változásokra" (1789; "On the Changes in France"), a vigorous warning to all tyrants "to cast their watchful eyes on Paris."

Sentimentalism found its exponents in József Kármán and Gábor Dayka. Kármán's only work of importance, *Fanni hagyományai* (1794; "The Memoirs of Fanny"), is a novel of sentiment written in the form of letters and diary entries. Very much on the lines of Goethe's *Werther*, the work nevertheless marks an important step in the history of the Hungarian novel. Dayka, who was a poet, died too young for the full measure of his talent to be realized.

The first important lyric poet since Bálint Balassi was Mihály Csokonai Vitéz, who continued the purely Hungarian poetical tradition. His many songs to a woman named Lilla are a happy blend of playful grace and subtle thoughts. The influence of Rousseau is very noticeable in some of his longer philosophical poems. Alexander Pope's *Rape of the Lock* served as a source of inspiration for Csokonai's comic epic *Dorottya* (1804), but Csokonai's poem is original and his context very Hungarian. The language of the poem is vigorous, even vulgar, and the plot is full of hilariously comic situations.

The place of Sándor Kisfaludy in Hungarian literature is secured by his first work, *Kesergő szerelem* (1801; "Bitter Love"), a lyric cycle depending on a very thin narrative thread. Writing in an elaborate verse form of 12 lines, called the Himfy verse, which he devised himself, Kisfaludy displayed great ingenuity in finding new variations on the theme of unhappy love.

Ferenc Kazinczy, a mediocre poet but an influential man of letters, was the pivot of literary life for about 40 years. For his involvement in the conspiracy of Martinovics he paid with six years' imprisonment. He wanted a literature refined and limpid, neither baroque nor popular, and his

Ferenc Kazinczy

interest was focused on style. He became the head of the *neologi,* or linguistic innovators, who tried to renew and enrich the Hungarian language so that it could express the most elaborate concepts. The success of the language reform was due, to a large extent, to Kazinczy's efforts.

THE 19TH CENTURY

Romanticism. The literary revival initiated by Kazinczy continued after his death. The literary leadership of Hungary at the beginning of the 19th century was assumed by Károly Kisfaludy when, in 1822, he founded a literary magazine, *Aurora,* to which all the important writers of the period contributed. He was also the first representative of Romanticism and the first playwright to achieve popular success.

While Kisfaludy's tragedies were applauded all over the country, *Bánk bán* (the *bán* was a high Hungarian dignitary), one of Hungary's best tragedies, by József Katona, was published in 1821 but, for the time being, was overlooked. Set in the 13th century and written in vigorous prose, the play was a masterful combination of national and individual conflicts, and one of its characters, Tiborc, a poor peasant, has remained ever since a symbol of the oppressed.

Ferenc Kölcsey was a deputy in the Hungarian parliament and a brilliant orator; his literary criticism was of a high standard, though unduly severe. His later poems, which were grave but vigorous in thought and expression, often dealt with national problems; his impressive "Hymnusz" (1823) became the Hungarian national anthem. After Kisfaludy's death, Mihály Vörösmarty became a central figure in literary life, producing writings of value in every genre. In particular he succeeded with a long epic poem *Zalán futása* (1825; "The Flight of Zalán"), written in a Romantic vein but expressing a concern for contemporary problems. This concern is evident also in many of his best lyric poems and even in his symbolic fairy play *Csongor és Tünde* (1831; "Csongor and Tünde").

Hungarian novels

In Hungarian literature, poetry was far ahead of drama, and the novel seemed slow in taking root. Miklós Jósika, a disciple of Sir Walter Scott, was the first successful novelist. His first and best work, the historical novel *Abafi* (1836), marked a turning point for the genre. József Eötvös, who after the 1848 revolution became a political theorist, produced two of the best novels in 19th-century Hungarian literature—*A falu jegyzője* (1845; *The Village Notary*), a portrait of feudal life in his own time, and *Magyarország 1514-ben* (1847; "Hungary in 1514"), about György Dózsa's peasants' revolt. They possessed exceptional qualities of characterization, both of individuals and of periods, and were political manifestos in support of the oppressed and against the appalling injustices that led to revolutions—of which Eötvös nevertheless disapproved.

The folk song and ballad collections of János Erdélyi and János Kriza exerted an influence on the further development of Hungarian poetry. "Popular poetry is the only real poetry" was the opinion of Sándor Petőfi, one of the greatest Hungarian poets, whose best poems rank among the masterpieces of world literature. He was an innovator and made a break with conventional subjects and poetic language. His poems are striking in immediacy of perception and directness of language and cover a vast range of subjects. The fervour of his patriotic poems inspired the revolution of 1848. Petőfi's many songs are enchanting in their simplicity, and in this genre he remained unsurpassed.

János Arany shared Petőfi's conviction of the value of popular poetry, but his approach was different, for his subjects were often taken from history and showed deep understanding of the human mind. He had the assurance of one who knew that what he wrote was the language of the people, lifted to a degree never surpassed in Hungarian. His ballads, often romantic, had vigour, conciseness, and uncommon evocative power. His great narrative poems, the *Toldi* trilogy (1847–79) and *Buda halála* (1864; *The Death of King Buda*), reflected eternal human problems; Arany's philosophy appeared through his characters and not in lengthy digressions and was accompanied by subtle humour.

Writers of the late 19th century. The peaks of poetry reached by Petőfi and Arany remained inaccessible to other poets during the rest of the 19th century. Hungary, after being defeated in the war of independence of 1848–49, was ruled from Vienna until 1867. External political pressures on Austria and the willingness of Hungarian society to end passive resistance made possible the Settlement (or Compromise) of 1867, which created the Austro-Hungarian Monarchy. The post-1867 industrial boom and Hungary's fast technological and commercial development produced a mood of complacency that was first broken by László Arany (son of János), whose ironic novel in verse, *A délibábok hőse* (1873; "The Hero of the Mirages"), is representative of the mood of disillusionment. Another poet, János Vajda, bridged the gap between the romantic populism of Petőfi and fin-de-siècle decadence: a gloomy visionary, with equal propensity for self-pity and self-aggrandizement, he was nevertheless an important innovator in the field of metaphor and poetic imagery.

The national theatre

In 1837 a national theatre was established to produce works of merit, but, with few exceptions, the standard of plays was low. Ede Szigligeti, a prolific playwright, wrote entertaining comedies and created a special genre of plays, the *népszinmü,* that give an idealized picture of village life but also contain a measure of social criticism. Very different were the plays of Imre Madách, whose masterpiece *Az ember tragédiája* (1861; *The Tragedy of Man*) dealt with universal human problems. This poetic drama followed man's destiny from creation through stages of history into a future of a phalanstery (a Utopian commune) and the ultimate extinction of life. The play was first staged in 1883 and remains a favourite with the Hungarian public.

The first outstanding novelist, Zsigmond Kemény, displayed, in such novels as *Zord idő* (1862; "Grim Times"), *A rajongók* (1858–59; "The Fanatics"), and *Férj és nő* (1852; "Husband and Wife"), a masterly skill in psychological analysis. His characters' own deeds determined their gloomy ends. Analysis often took the place of action in Kemény's novels, which were therefore difficult to read and not popular. On the other hand, Mór Jókai was a popular Hungarian novelist, an exceptional storyteller able to evoke any epoch and any milieu. His characters were idealized, and his descriptions tended to be brilliant rather than accurate. Among his numerous works (he published more than 200 books in his lifetime) were historical novels on problems of contemporary society. *Az arany ember* (1873; "That Golden Man"; Eng. trans., *Timár's Two Worlds*) is one of his best novels. Kálmán Mikszáth was also popular; he recorded with keen observation and sly humour the shortcomings of society but, although a politician and a member of parliament, was little concerned with improvement. Though the principal works of Géza Gárdonyi were published early in the 20th century, they belonged to the 19th century. *Egri csillagok* (1901; "The Stars of Eger") and *A láthatatlan ember* (1902; "The Invisible Man"; Eng. trans., *Slave of the Huns*) were well constructed.

During the 19th century, literary life in Hungary became organized to some extent: the Kisfaludy and Petőfi societies, founded in 1836 and 1876, respectively, were particularly influential, though the authority of the Hungarian Academy, founded in 1825, remained unchallenged. The principal critic of the second half of the century was Pál Gyulai. Literary criticism and the history of Hungarian literature attracted some of the best minds, including Jenő Péterfy and Frigyes Riedl.

The last third of the 19th century in Hungary was an era of literary decline in which writers based their work on social and political ideals that were becoming sterile. The great majority of Hungarian writers came from the nobility and lived as part of the middle class; only at the end of the century did lower-middle-class writers come to the fore. The periodical *A hét* ("The Week"), founded in 1890 by József Kiss, became the organ of a number of gifted writers, including Zoltán Ambrus and Sándor Bródy.

THE 20TH CENTURY

Early years. The year 1906, when Endre Ady burst upon the literary scene with his *Uj versek* ("New Poems"),

The influence of the periodical *Nyugat*

marked a turning point. In matters of style Ady was influenced by the French Symbolists, but in content he was concerned with radical political ideas. He rejuvenated the language of Hungarian poetry, introducing new themes and powerful new imagery. His rise was helped by the periodical *Nyugat* ("The West"), which was launched in 1908 under the editorship of Hugo Ignotus, Miksa Fenyő, and Ernő Osvát. Among poets associated with *Nyugat* were Mihály Babits, an excellent translator of foreign poetry who became editor in 1929; Dezső Kosztolányi, who wrote with empathy on childhood and death and whose novels and short stories established high standards in narrative prose; and Árpád Tóth and Gyula Juhász, who voiced the distress of the poor and the oppressed in society. A fifth poet, Milán Füst, wrote little, but the dramatic metaphors and sonorous language of the work he did produce made his a lasting influence. In addition to his poetry he wrote an outstanding novel, *A feleségem története* (1942; "The Story of My Wife").

The prose writers of *Nyugat* included Zsigmond Móricz, whose tales of provincial life portrayed peasants and gentry; Margit Kaffka, the first major woman writer in Hungary; and Gyula Krúdy, who created a nostalgic dreamworld with his stream-of-consciousness technique.

Writers not connected with *Nyugat* included the versatile Ferenc Molnár, who, after a promising start as a writer of fiction, began to write cleverly constructed social comedies. A conservative-nationalist group of writers was highly influential before 1918; its principal figure was Ferenc Herczeg, an author of novels and plays. During World War I and the years of revolution that followed, two authors emerged to challenge both the old establishment and *Nyugat*. These were Lajos Kassák, the first significant poet of the Hungarian avant-garde, who also wrote a remarkable autobiography depicting working-class life at the beginning of the century; and Dezső Szabó, whose large, uneven expressionistic novel *Az elsodort falu* (1919; "The Village That Was Swept Away") combined antiwar sentiment with a romantic cult of the peasantry. First embraced and then rejected by the post-1919 counterrevolution, Szabó is best remembered as a witty though venomous pamphleteer.

The interwar period. The interwar period saw a flowering of Hungarian letters. Although the influence of *Nyugat* diminished, neither the populist *Válasz* ("The Response") nor the left-wing *Szép szó* ("Fine Word") could quite supplant it. The leading poet of the 1920s was Lőrinc Szabó, a master of poetic technique and fine observation, whereas the 1930s were dominated by Attila József, whose experience of alienation and Socialist ideas were expressed in great poetic tableaux and in poems probing the subconscious, and by Gyula Illyés, who found inspiration in the life of the peasantry. The poetry of Miklós Radnóti reached a tragic climax in the serene and polished poems he wrote in the last years of his life.

In Hungary, as elsewhere, the novel became the principal form of literary expression. While Sándor Márai and Lajos Zilahy depicted the life of the bourgeoisie, János Kodolányi, László Németh, and Zsigmond Remenyik exposed the conflicts of the individual with society (often against a background of injustice and misery). Áron Tamási wrote beautifully stylized novels on the life of the Szeklers, an ethnic group of Transylvania. Tibor Déry, whose chief work was published only after 1945, wrote realistic novels and a challenging autobiography. The foremost essayist of the period was László Németh, whose *A minőség forradalma* (1940; "The Revolution of Quality") remained a seminal influence for many years to come. Other essayists and literary historians active in this era included a particularly brilliant writer, Antal Szerb, and Gábor Halász (both died in forced labour camps in 1945) and László Cs. Szabó.

Writing after 1945. The period since 1945, though officially designated one of "Socialist transformation," has seen but little change in writers' traditional orientations and preoccupations. During the first decade, particularly the years 1948–53, many writers were forced into silence by the regime's attempts to introduce Socialist Realism as the only correct style and creative method. After the failure of the 1956 uprising a number of writers were imprisoned, but by the mid-1960s most efforts to enforce ideological purity in the arts were abandoned. Since then there has been comparatively little official intervention in Hungarian literature and the margin of free experimentation has grown. This allowed writers such as Géza Ottlik, Miklós Mészöly, and István Örkény to publish work that showed ways in which the technique of modern fiction could be applied in Hungary. Among the best new authors were György Konrád and Péter Esterházy. Konrád's novels *A látogató* (1969; *The Case Worker*), *A városalapító* (1977; *The City Builder*), and the unofficially published *A cinkos* (1982; *The Loser*) achieved great impact with their dense, poetically structured style and analytical probing into the world of the social caseworker, the planner of new society, and the mental institution. Esterházy's most successful novel to date, *Termelési regény (kisssregény)* (1979; "Production Novel (a Ssshort [sic] Novel)"), is a grotesque and refreshingly irreverent survey of Hungarian life and society.

Among the adherents of realistic fiction, József Lengyel, who died in 1975, occupied a special place. In his stories (which could not be published until the loosening of restrictions in the early 1960s) he gave a moving testimony of human suffering in Soviet labour camps.

The best poetry was written by Sándor Weöres, whose poetic span ranges from Eastern philosophy to delightful children's verses, and János Pilinszky, an Existentialist Catholic whose most memorable poems deal with the experience of what he called the "universe of camps" produced by World War II. Other noteworthy poets included the urbane László Kálnoky and István Vas, and Ferenc Juhász and László Nagy, two poets of peasant origin whose work grew out of native tradition to express universal rites and myths of mankind such as marriage, the struggle among generations for power, and cosmic destruction.

Frontier changes since World War I have placed substantial Hungarian minorities in countries outside Hungary, especially in neighbouring Czechoslovakia, Yugoslavia, and Romania. In Romania, for example, where approximately 2,000,000 Hungarians live, the best known Hungarian writer is the playwright and novelist András Sütő. There also has been a large diaspora in the West, where, apart from Márai, the versatile modernist Győző Határ and the post-Romantic poet György Faludy have the largest following. The Munich-based cultural review *Uj látóhatár* ("New Horizon") has enjoyed the longest uninterrupted existence among Hungarian periodicals inside or outside of Hungary.

BIBLIOGRAPHY. ALBERT TEZLA, *An Introductory Bibliography to the Study of Hungarian Literature* (1964), and *Hungarian Authors: A Bibliographical Handbook* (1970), are annotated sources on items available from major U.S. and European book collections. ANTAL SZERB, *Magyar irodalomtörténet,* 6th ed. (1978), is an informative history, an earlier edition of which is available also in a German translation: *Ungarische Literaturgeschichte,* 2 vol. (1975). An exhaustive historical coverage is provided in ISTVÁN SÖTÉR (ed.), *A magyar irodalom története,* 6 vol. (1964–66). BÉLA POMOGÁTS, *Az újabb magyar irodalom, 1945–1981* (1982), is an informative critical study of 20th-century literature. In English, see D. MERVYN JONES, *Five Hungarian Writers* (1966), which gives a detailed analysis of the works of the writers Miklós Zrínyi, Kelemen Mikes, Mihály Vörösmarty, József Eötvös, and Sándor Petőfi. TIBOR KLANICZAY (ed.), *A History of Hungarian Literature* (1982; originally published in Hungarian, 1982), provides broad coverage. LÓRÁNT CZIGÁNY, *The Oxford History of Hungarian Literature from the Earliest Times to the Present* (1984), is a comprehensive chronologically arranged work, with emphasis on the 19th and 20th centuries.

(T.Kl./D.Si./G.G.)

Hungary

Located in the heart of Europe and occupying part of the Carpathian Basin, the nation of Hungary has ethnic and linguistic roots that reach far back into the past. Hungary (in full, Republic of Hungary; Hungarian: Magyar Köztársaság) is a member of the Warsaw Pact and the Council for Mutual Economic Assistance (Comecon). It has an area of 35,919 square miles (93,031 square kilometres).

Hungary is bordered by Czechoslovakia to the north, the Soviet Union and Romania to the east, Yugoslavia to the south, and Austria to the west. Budapest, the capital city, which dominates much of national life, is situated a few miles downstream from the Danube Bend, where the Danube (Hungarian: Duna) abruptly changes course from an easterly to a southerly direction.

This article is divided into the following sections:

Physical and human geography 693
- The land 693
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 695
  - Ethnic and religious structure
  - Demographic trends
- The economy 696
  - Resources
  - Agriculture
  - Industry
  - Finance and trade
  - Transportation
- Administration and social conditions 698
  - Government
  - Justice
  - Armed forces
  - Education
  - Health and welfare
  - Housing
- Cultural life 699
  - The cultural milieu
  - Cultural institutions

History 700
- The kingdom to 1526 700
  - The Árpáds
  - Hungary under foreign kings
- The period of partition 703
  - Royal Hungary and the rise of Transylvania
  - War and liberation
- Habsburg rule, 1699–1918 705
  - Habsburg rule to 1867
  - The Dual Monarchy, 1867–1918
- Revolution, counterrevolution, and the Regency, 1918–45 708
  - The Regency, 1920–45
  - Financial crisis: the rise of right radicalism
  - War and renewed defeat
- The People's Republic 710
  - The communist regime
  - The revolution of 1956
  - The Kádár regime
  - Reforms of the late 1980s

Bibliography 714

Physical and human geography

THE LAND

The Great and Little Alfolds

Relief. Dominating the relief are the great lowland expanses that make up the core of Hungary. The Little Alfold (Little Hungarian Plain, or Kis Alföld) lies in the northwest, fringed on the west by the easternmost extension of the sub-Alps along the border with Austria and bounded on the north by the Danube. The Little Alfold is separated from the Great Alfold (Great Hungarian Plain, or Nagy Magyar Alföld) by a low mountain system extending across the country from southwest to northeast for a distance of 250 miles (400 kilometres). This system, which forms the backbone of the country, is made up of Transdanubia (Dunántúl) and the Northern Mountains, separated by the Visegrád gorge of the Danube. The former is dominated by the Bakony Mountains, which consist of dolomite and limestone plateaus at heights between 1,300 and 2,300 feet (400 and 700 metres) above sea level interspersed with volcanic peaks; the latter, which consist of volcanic rocks, comprise the Mátra Mountains in the north, reaching a height of 3,330 feet (1,015 metres) at Kékes, the highest peak in Hungary. Regions of hills reaching altitudes of 800 to 1,000 feet lie on either side of the mountain backbone, while to the south and west of Lake Balaton is an independent upland region of more subdued, loess-covered topography. The Great Alfold covers most of central and southeastern Hungary. Like its northwestern counterpart, it is a basinlike structure filled with fluvial and windblown deposits. Four types of surface may be distinguished: floodplains, composed of river alluvium; alluvial fans, wedge-shaped features deposited at the breaks of slopes where the rivers emerge from the mountain rim; alluvial fans overlain by sand dunes; and plains buried under loess, a windblown deposit derived from the continental interior. These lowlands range in height from about 260 to 660 feet above sea level, with the lowest point at 256 feet, on the southern edge of Szeged, along the Tisza River.

Drainage and soils. Hungary lies within the drainage basin of the Danube. The Danube and two of its tributaries, the Rába and the Drava, are of Alpine origin, while the Tisza and its tributaries, which drain much of eastern Hungary, rise in the Carpathians.

The Danube floods regularly twice a year, first in early spring and again in early summer. During these phases, discharge is up to 10 times greater than that recorded during the low-water periods of autumn and winter. Where it flows through Hungary, the Tisza is of floodplain character, and large meanders and oxbow lakes marking former channels are typical features. At Szolnok, peak discharges 50 times greater than average have been recorded. In the past, devastating floods occurred on the Danube, the Tisza, and their tributaries. About 2,500 miles of levees have been built to protect against floods. The arid climate of the central and eastern areas of the Great Alfold has necessitated the construction of large-scale irrigation systems, mostly along the Tisza River.

There are few lakes in Hungary, and these, for the most part, are small. Lake Balaton, however, is the largest freshwater lake in central Europe; Velence Lake and Lake Fertő (in Austria, Lake Neusiedler) lie on the Austrian border.

In the forest zones, gray-brown podzolic (leached) and brown forest soils predominate, while under the forest-steppe rich black-earth, or chernozem, soil has evolved, although sand dunes and dispersed alkali soils are also characteristic.

Temperatures

Climate. Because of its situation within the Carpathian Basin, Hungary has a moderately dry continental climate. The mean annual temperature is about 50° F (10° C), ranging from 32° to 25° F (0° to −4° C) in January to July temperatures of 64° to 73° F (18° to 23° C). A



MAP INDEX

Political subdivisions

Bács-Kiskun . . . . . 46 30 N 19 25 E
Baranya . . . . . 46 05 N 18 15 E
Békés . . . . . 46 45 N 21 00 E
Borsod-Abaúj-Zemplén . . . . . 48 15 N 21 00 E
Budapest . . . . . 47 30 N 19 05 E
Csongrád . . . . . 46 25 N 20 15 E
Fejér . . . . . 47 10 N 18 35 E
Győr-Sopron . . . . . 47 40 N 17 15 E
Hajdú-Bihar . . . . . 47 25 N 21 30 E
Heves . . . . . 47 50 N 20 15 E
Komárom . . . . . 47 37 N 18 20 E
Nógrád . . . . . 48 00 N 19 35 E
Pest . . . . . 47 25 N 19 20 E
Somogy . . . . . 46 25 N 17 35 E
Szabolcs-Szatmár . 48 00 N 22 10 E
Szolnok . . . . . 47 15 N 20 30 E
Tolna . . . . . 46 30 N 18 35 E
Vas . . . . . 47 10 N 16 45 E
Veszprém . . . . . 47 10 N 17 40 E
Zala . . . . . 46 40 N 16 50 E

Cities and towns

Abony . . . . . 47 11 N 20 00 E
Ajka . . . . . 47 06 N 17 34 E
Badacsonytomaj . . 46 48 N 17 31 E
Baja . . . . . 46 11 N 18 58 E
Balassagyarmat . . 48 05 N 19 18 E
Balatonfüred . . . . . 46 57 N 17 53 E
Balmazújváros . . . . 47 37 N 21 21 E
Barcs . . . . . 45 58 N 17 28 E
Bátonyterenye . . . . 47 58 N 19 50 E
Békés . . . . . 46 46 N 21 08 E
Békéscsaba . . . . . 46 41 N 21 06 E
Berettyóújfalu . . . . 47 13 N 21 33 E
Bicske . . . . . 47 29 N 18 38 E
Bonyhád . . . . . 46 18 N 18 32 E
Budakeszi . . . . . 47 31 N 18 56 E
Budaörs . . . . . 47 27 N 18 58 E
Budapest . . . . . 47 30 N 19 05 E
Cegléd . . . . . 47 10 N 19 48 E
Celldömölk . . . . . 47 15 N 17 09 E
Csongrád . . . . . 46 42 N 20 09 E
Csorna . . . . . 47 37 N 17 15 E
Dabas . . . . . 47 11 N 19 19 E
Debrecen . . . . . 47 32 N 21 38 E
Dombóvár . . . . . 46 23 N 18 07 E
Dorog . . . . . 47 43 N 18 44 E
Dunaharaszti . . . . . 47 21 N 19 05 E
Dunakeszi . . . . . 47 38 N 19 08 E
Dunaújváros (Sztálinváros) . . . . 46 59 N 18 56 E
Eger . . . . . 47 54 N 20 23 E
Esztergom . . . . . 47 48 N 18 45 E
Fertőd (Eszterháza) . . . . 47 37 N 16 52 E
Fót . . . . . 47 37 N 19 12 E
Göd . . . . . 47 42 N 19 08 E
Gödöllő . . . . . 47 36 N 19 22 E
Gyál . . . . . 47 23 N 19 14 E
Gyomaendrőd . . . . 46 56 N 20 50 E
Gyömrő . . . . . 47 25 N 19 24 E
Gyöngyös . . . . . 47 47 N 19 56 E
Győr . . . . . 47 41 N 17 38 E
Gyula . . . . . 46 39 N 21 17 E
Hajdúböszörmény . 47 40 N 21 31 E
Hajdúhadház-téglás . . . . . 47 41 N 21 40 E
Hajdúnánás . . . . . 47 51 N 21 26 E
Hajduszoboszló . . . 47 27 N 21 24 E
Hatvan . . . . . 47 40 N 19 41 E
Heves . . . . . 47 36 N 20 17 E
Hévíz . . . . . 46 47 N 17 11 E
Hódmezővásár-hely . . . . . 46 25 N 20 20 E
Jánoshalma . . . . . 46 18 N 19 20 E
Jászapáti . . . . . 47 31 N 20 09 E
Jászberény . . . . . 47 30 N 19 55 E
Kalocsa . . . . . 46 32 N 19 00 E
Kaposvár . . . . . 46 22 N 17 48 E
Kapuvár . . . . . 47 36 N 17 02 E
Karcag . . . . . 47 19 N 20 56 E
Kazincbarcika . . . . 48 15 N 20 38 E
Kecskemét . . . . . 46 54 N 19 42 E
Kerepestarcsa . . . . 47 32 N 19 16 E
Keszthely . . . . . 46 46 N 17 15 E
Kiskőrös . . . . . 46 37 N 19 18 E
Kiskunfélegyháza . 46 43 N 19 51 E
Kiskunhalas . . . . . 46 26 N 19 30 E
Kisújszállás . . . . . 47 13 N 20 46 E
Kisvárda . . . . . 48 13 N 22 05 E
Komárom . . . . . 47 44 N 18 07 E
Komló . . . . . 46 12 N 18 16 E
Körmend . . . . . 47 01 N 16 36 E
Kőszeg . . . . . 47 23 N 16 33 E
Kővágószőllős . . . . 46 05 N 18 08 E
Leninváros . . . . . 47 56 N 21 05 E
Lenti . . . . . 46 37 N 16 33 E
Makó . . . . . 46 13 N 20 29 E
Marcali . . . . . 46 35 N 17 25 E
Mátészalka . . . . . 47 57 N 22 20 E
Mezőberény . . . . . 46 49 N 21 02 E
Mezőkövesd . . . . . 47 49 N 20 35 E
Mezőtúr . . . . . 47 00 N 20 38 E
Miskolc . . . . . 48 06 N 20 47 E
Mohács . . . . . 45 59 N 18 42 E
Monor . . . . . 47 21 N 19 27 E
Mór . . . . . 47 23 N 18 12 E
Mosonmagyaróvár . 47 52 N 17 17 E
Nagyatád . . . . . 46 13 N 17 22 E
Nagykanizsa . . . . . 46 27 N 16 59 E
Nagykáta . . . . . 47 25 N 19 45 E
Nagykőrös . . . . . 47 02 N 19 47 E
Nyírbátor . . . . . 47 50 N 22 08 E
Nyíregyháza . . . . . 47 57 N 21 43 E
Orosháza . . . . . 46 34 N 20 40 E
Oroszlány . . . . . 47 29 N 18 19 E
Ózd . . . . . 48 13 N 20 18 E
Paks . . . . . 46 38 N 18 52 E
Pápa . . . . . 47 20 N 17 28 E
Pécs . . . . . 46 05 N 18 14 E
Pomáz . . . . . 47 39 N 19 02 E
Püspökladány . . . . 47 19 N 21 07 E
Rudabánya . . . . . 48 23 N 20 38 E
Sajószentpéter . . . 48 13 N 20 43 E
Salgótarján . . . . . 48 07 N 19 49 E
Sárbogárd . . . . . 46 53 N 18 38 E
Sarkad . . . . . 46 45 N 21 23 E
Sárospatak . . . . . 48 19 N 21 35 E
Sárvár . . . . . 47 15 N 16 56 E
Sátoraljaújhely . . . . 48 24 N 21 40 E
Siklós . . . . . 45 51 N 18 18 E
Siófok . . . . . 46 54 N 18 03 E
Sopron . . . . . 47 41 N 16 36 E
Szarvas . . . . . 46 52 N 20 33 E
Szeged . . . . . 46 15 N 20 10 E
Szeghalom . . . . . 47 02 N 21 10 E
Székesfehérvár . . . 47 12 N 18 25 E
Szekszárd . . . . . 46 21 N 18 43 E
Szentes . . . . . 46 39 N 20 16 E
Szigetszentmiklós . 47 21 N 19 03 E
Szigetvár . . . . . 46 03 N 17 48 E
Szolnok . . . . . 47 11 N 20 12 E
Szombathely . . . . . 47 14 N 16 37 E
Sztálinváros, see Dunaújváros
Tamási . . . . . 46 38 N 18 17 E
Tapolca . . . . . 46 53 N 17 26 E
Tata . . . . . 47 39 N 18 19 E
Tatabánya . . . . . 47 34 N 18 25 E
Tiszaföldvár . . . . . 46 59 N 20 15 E
Tiszafüred . . . . . 47 37 N 20 46 E
Tiszavasvári . . . . . 47 58 N 21 21 E
Tokaj . . . . . 48 07 N 21 25 E
Törökszentmiklós . . 47 11 N 20 25 E
Túrkeve . . . . . 47 06 N 20 45 E
Újfehértó . . . . . 47 48 N 21 41 E

Vác . . . . . 47 47 N 19 08 E
Várpalota . . . . . 47 12 N 18 08 E
Vecsés . . . . . 47 24 N 19 17 E
Veszprém . . . . . 47 06 N 17 55 E
Visegrád . . . . . 47 47 N 18 59 E
Zalaegerszeg . . . . 46 50 N 16 51 E

Physical features and points of interest
Bakony Mountains . . . . . 47 15 N 17 50 E
Balaton, Lake . . . . 46 50 N 17 45 E
Baranya, see Mecsek
Bükk Mountains . . 48 05 N 20 30 E
Bükk National Park . . . . . 48 06 N 20 30 E
Csóványos, Mount . . . . . 47 57 N 18 57 E
Danube (Duna), *river* . . . . . 47 00 N 19 00 E
Drava (Dráva), *river* . . . . . 45 33 N 18 55 E
Dunántúl, see Transdanubia
Dunavölgyi Canal . . . . . 46 12 N 18 56 E
Eger, *river* . . . . . 47 38 N 20 39 E
Fertő (Neusiedler), Lake . . . . . 47 50 N 16 45 E
Great Alfold (Great Hungarian Plain, or Nagy Magyar Alföld), *lowland* . . . . . 47 00 N 20 00 E
Hajdúság, *region* . . 47 35 N 21 30 E
Hortobágy, *region* . . . . . 47 35 N 21 05 E
Hortobágy National Park . . . . 47 34 N 21 09 E
Istállóskő, Mount . . 48 04 N 20 26 E
Kapos, *river* . . . . . 46 44 N 18 29 E
Kékes, Mount . . . . 47 52 N 20 01 E
Keleti Main Canal . . . . . 48 01 N 21 20 E
Kis Alföld, see Little Alfold
Kiskunság, *region* . . . . . 46 35 N 19 15 E
Kiskunság National Park . . . . 46 40 N 19 30 E
Kőris, Mount . . . . . 47 18 N 17 45 E
Körös, *river* . . . . . 46 43 N 20 12 E
Little Alfold (Little Hungarian Plain, or Kis Alföld), *lowland* . . . . . 47 30 N 17 00 E
Marcal, *river* . . . . . 47 38 N 17 32 E
Maros (Mureşul), *river* . . . . . 46 15 N 20 12 E
Mátra Mountains . . 47 53 N 19 57 E
Mecsek (Baranya) Mountains . . . . . 46 10 N 18 18 E
Mura (Mur), *river* . . . . . 46 18 N 16 55 E
Nagy Magyar Alföld, see Great Alfold
Neusiedler, see Fertő
Nógrádi, *river* . . . . 47 40 N 19 41 E
Northern Mountains, *region* . . . . . 48 00 N 20 30 E
Nyírség, *region* . . . 47 50 N 21 55 E
Nyugati Main Canal . . . . . 47 57 N 21 19 E
Rába, *river* . . . . . 47 41 N 17 38 E
Sió, *river* . . . . . 46 22 N 18 48 E
Tisza (Tisa), *river* . . . . . 45 15 N 20 17 E
Tiszántúl, see Transtisza
Transdanubia (Dunántúl), *region* . . . . . 47 00 N 18 00 E
Transtisza (Tiszántúl), *region* . . . . . 47 00 N 21 00 E
Velence (Velencei), Lake . . 47 13 N 18 36 E

summer temperature as high as 109° F (43° C) has been recorded, while the absolute winter minimum is −29° F (−34° C). In the lowlands, precipitation generally ranges from 20 to 24 inches (550 to 600 millimetres), rising to 24 to 31 inches over higher elevations. The central and eastern areas of the Great Alfold are the driest and the southwestern uplands the wettest part of the country. As much as two-thirds of annual precipitation falls during the growing season.

Plant and animal life. Human activities over the ages have largely destroyed the natural vegetation of Hungary. Some three-fifths of the land is regularly cultivated; the remainder comprises meadows and rough pasture and forest and woodland, with a small percentage used for nonagricultural purposes. No part of the country is of sufficient altitude to support natural coniferous forest, beech being the climax species at the highest elevations and oak woodland alternating with scrubby grassland being the climax at lower altitudes in the upland regions.

Deer and wild pigs are abundant in the forests at higher altitudes, while the lowlands have an animal population of rodents, hares, partridge, and pheasant. The once numerous varieties of marsh waterfowl survive only in nature reserves. There are diverse species of freshwater fish, including pike, bream, and pike perch. Significant water and air pollution occurs in some of the industrial regions of the country.

Settlement patterns. *Traditional regions.* The Great Alfold is the largest region of the country. It is divided into two parts: Kiskunság, the area lying between the Danube and Tisza rivers, and Transtisza (Tiszántúl), the region east of the Tisza. The former consists primarily of a mosaic of small landscape elements—sand dunes, loess plains, and floodplains. Kecskemét is the market centre for the region, which is also noted for its isolated farmsteads, known as *tanyák*. Several interesting groups live there, including the people of Kalocsa and the Matyó, who live in the northern part of the plain around Mezőkövesd and are noted for hand embroidery and multicoloured costumes. In the generally homogeneous flat plain of the Transtisza region, only the Nyírség presents any form of topographical contrast. Closely connected with this latter area are the Hajdúság and the Hortobágy, and all three areas look to Debrecen, the largest city of the plain. The Hortobágy presents an interesting survival of the steppe life of earlier times, since the original Hungarian cattle, horse, and sheep breeds have been preserved there as part of the national heritage.

The Little Alfold, the second major natural region, is situated in the northwest and is traversed by the Danube and Rába rivers and their tributaries. It is more favourably endowed naturally than the Great Alfold; both agriculture and industry are more advanced. Győr, known for its Baroque architecture, is the major city of the region.

The third major region is Transdanubia, which embraces all of the country west of the Danube. It is a rolling upland broken by the Bakony and Mecsek ridges. Lake Balaton is a leading resort area. To the south of the lake are the hills of Somogy, Tolna, and Baranya, where the people view Pécs, a mining and industrial city, as the economic and cultural centre. Also found in Transdanubia are the Bakony Mountains, whose isolation, densely forested ridges, small closed basins, and medieval fortresses and monasteries have afforded protection to the local inhabitants over the course of many stormy centuries. Although modern industrial towns, drawing on the bauxite, manganese, and brown coal resources of the area, have sprung up, the cultural centre of the area is the historic city of Veszprém. In the southern part of the region, west of Lake Balaton, are health resorts and centres of wine production, notably Keszthely, Hévíz, Badacsony, and Balatonfüred.

The Northern Mountains, the fourth major geographic region of the country, contain two important industrial areas, the Nógrád and Borsod basins. Agriculture is important, especially viticulture, including the well-known Tokay and Eger vineyards. Tourism is well-developed, and numerous spas and recreation centres are located there. Miskolc is the main economic centre for the region.

Urban settlement. More than half the population is urban, but the majority of towns in Hungary have populations of less than 40,000. They were functionally, until the late 20th century, vastly overgrown villages rather than towns. More than one-third of the urban population lives within the Budapest metropolitan area.

Urban Hungary is dominated by Budapest, which is several times the size of other major cities. It has the largest industrial work force in the country. The major provincial centres are Miskolc, Debrecen, Szeged, Pécs, and Győr, each of which has a population exceeding 100,000; an economic, cultural, and administrative hinterland that reaches deep into the surrounding countryside; and an expanding industrial capacity. Below the provincial centres in the hierarchy are the traditional market towns, such as Kecskemét, Székesfehérvár, Nyíregyháza, Szombathely, and Szolnok, with new suburbs added to medieval or Baroque town centres. Also worthy of note are the predominantly industrial towns located close to the mineral resources of the Northern Mountains, which, from small beginnings in the late 19th century, have developed into major industrial centres. They include Tatabánya, Salgótarján, and Ózd. In addition, a number of industrial towns were created in the late 20th century on green-field sites as part of deliberate planning policy. These include the metallurgical centre of Dunaújváros on the Danube and the chemical centre of Kazincbarcika in eastern Hungary.

Rural settlement. The distribution of rural population varies widely from one part of the country to another. For historical reasons connected with resettlement following the Turkish occupation in the 16th century, the compact villages of the Great Alfold are small in number but large in size. By comparison, rural settlement in Transdanubia and in the Northern Mountains takes the form of many small nucleated and linear villages. The *tanyák* tend to be concentrated in the Great Alfold.

Rural population distribution

THE PEOPLE

Ethnic and religious structure. The mother tongue of almost the entire population is Hungarian, or Magyar,

which belongs to the Finno-Ugric group of the Uralic linguistic family and is related only to Finnish and Estonian among the languages of Europe. The Hungarian people, known in the vernacular as Magyars, migrated from the steppes of eastern Europe and settled in the Carpathian Basin during the 9th century AD. Today they are a homogeneous people, and subgroups within them are difficult to distinguish. Magyars constitute about 92 percent of Hungary's citizenry. Half of the remaining population are Gypsies, while the balance includes Germans, Slovaks, Serbians, Croatians, and Romanians.

Since 1949, Hungary has not published data on the population's religious affiliation. In that year, more than two-thirds of the people were Roman Catholic, and just under 22 percent were Calvinist. Lutherans constituted 5.2 percent and Uniates (Greek Catholics) 2.7 percent of the citizenry. There were also smaller numbers of Jews, Greek Orthodox, and Unitarians, constituting altogether 2 percent of the population.

Demographic trends. Paralleling the trends exhibited by most countries of western and central Europe, the rate of population growth in Hungary during the 20th century has been comparatively low. Official population policy, however, encourages a higher birth rate by making abortion more difficult to obtain, giving preference in housing allocation and house-building grants to families with children, safeguarding the employment rights of mothers, and providing generous child allowances. Such earlier gains, however, have been undermined by a deteriorating economy, a chronic shortage of low-income housing, and a new income-tax law that ignores the size of the family.

In the 1980s the crude death rate exceeded the birth rate, resulting in a negative rate of natural increase. Life tables suggest that male life expectancy at birth is now virtually static, while female longevity continues to increase (at an ever lower rate).

The proportion of the population over the age of 60 years has been rising steadily. The migration of young people from the countryside to the towns has meant that population aging has been more accentuated in rural than in urban areas. Since the late 19th century there has been a female surplus. The heavy male mortality during the two world wars strengthened the imbalance, and the female preponderance is especially evident among people over the age of 40. Migration again produces geographic disparities in the sex ratio; although the female surplus is greater in rural than in urban areas because more men than women migrate to the towns, it is most pronounced among the resident population of Budapest.

A surplus of women

After World War I some two-thirds of the former territory of Hungary was annexed by neighbouring states, and as a consequence more than three million Hungarians live outside the country in Romania, Czechoslovakia, Yugoslavia, and the Soviet Union. During the aftermath of the Hungarian Revolution of 1956, 200,000 Hungarians sought homes in other countries; of these, 20,000 later returned to Hungary.

THE ECONOMY

Hungary is a predominantly industrial country, with less than 20 percent of the economically active population involved in agriculture. The economy is divided into state, cooperative, and private sectors. Large industrial and agricultural enterprises are owned by the state. The cooperative sector is made up of farming, manufacturing, and artisans' cooperatives and a number of marketing and consumption cooperatives. The private sector of the economy comprises artisans, retail tradesmen, small farmers, and entrepreneurs. The majority of national income is generated by the state sector, followed by cooperatives and private initiative.

The New Economic Mechanism

In 1968 a reformed system of economic management, the New Economic Mechanism (NEM), was introduced, replacing the previous rigid system of central planning—the direction of the economy by means of centrally determined, obligatory targets for individual enterprises. The basic aim of the NEM was to improve economic efficiency, both in domestic production and in foreign trade activities. Achieving this involved the abolition of directive planning and an increase in the importance of prices in determining decisions about enterprises. Producers' prices were reformed to bring them more into line with production costs, and a system was devised to allow the payment of bonuses to workers and management in accordance with profits achieved rather than plan norms. Enterprises were allowed a degree of autonomy over future investment. They also handled their own foreign trade relations. Joint ventures with Western companies were permitted.

The reform was generally considered to have been successful. More balanced growth was maintained than in the earlier period, although controls on prices and profits were tightened again in 1973. This undermined the NEM and contributed to economic difficulties. In the late 1980s the economy was again in the process of transition from central planning to a more market-oriented system. Due to stagnation of the economy, increased inflationary pressures, and a considerable decline in earnings, the per capita indebtedness of the country skyrocketed in the 1980s and was the highest among eastern European countries. Gradual withdrawal of government subsidies for inefficient enterprises led to increased unemployment, and adjustment of consumer prices to the costs of production resulted in rising inflation and a general decline in living standards. Economic reforms encouraged further development of private enterprises and aimed at restoring balance in the country's international accounts, halting the rise of the national debt, and gradually reducing the budget deficit by restrictions on imports and internal consumption. A new bankruptcy law, which allows company managers or creditors to initiate bankruptcy proceedings through the courts, took effect in 1987. The state intervenes with financial assistance only if the closure of an enterprise would affect a large number of workers or if it would jeopardize national security.

The main sources of state finance are the taxes and other payments collected from nationalized and cooperative enterprises. Taxes paid directly by the population include an annual property tax, an income tax, a gift tax, and death duties. A value-added tax was introduced in 1988; essential food items, household fuel, medicines, and housing are exempt from this taxation.

Trade unions

Trade unions affiliated to the Council of Hungarian Trade Unions used to be controlled by the Communist Party, but the government has accepted a growing number of independent trade unions since 1988. The functions of trade unions in Hungary differ considerably from those in the West. On the one hand, they are concerned with such economic matters as wage levels and job protection. This role was increased within the framework of the NEM and the introduction of a modified labour code in which many of the restrictions on termination of employment and the obligation of the state to fix wages were abolished. Collective agreements between the enterprise management and the union have thus taken on greater significance. By contrast with usual practice in the West, unions are obliged to further agreed government policy. Disputes with management over collective agreements are settled through the courts, although use of the strike has been legalized. Enterprise managements are compelled by law to consult the unions about the living and working conditions of employees. Hungarian unions also serve an important social and cultural role. Union holiday homes are to be found in the principal resort areas of the country, providing inexpensive vacations for members. Unions also control the social and cultural funds accumulated by enterprises.

Resources. The major natural endowments of Hungary are fertile soils and a climate favourable for agriculture. The power resources are modest, and the only significant mineral wealth is bauxite and the manganese deposits associated with it in the Bakony Mountains. The production of iron ore has virtually ceased. The basic materials for the construction industry are mostly found within the country. Coal output is mainly of low-grade brown coal from mines in the Northern Mountains; black coal of good quality is mined only at Komló near Pécs.

In addition to the Transdanubian fields, which have been worked since the 1930s, extensive oil and natural

gas fields have been discovered and explored at several points of the Great Alfold. Natural gas is produced in the vicinity of Szeged. There are copper deposits at Recsk, but the mines remain undeveloped. Other minerals include mercury, lead, uranium, and molybdenum.

Agriculture. Before World War II the economy of Hungary was based primarily on agriculture, but the industrialization drive begun in 1950 fundamentally altered the balance. Whereas agriculture accounted for about half the national income prior to this industrialization, the proportion has fallen to one-fifth.

Hungary is virtually self-sufficient in food production and exports several agricultural products. Corn (maize), wheat, sugar beets, barley, and potatoes are the principal crops, but other vegetables and beef, pork, and poultry production are also important. About two-thirds of the total area of the country is under cultivation; orchards and vineyards are also notable.

Agriculture is organized into cooperative and state farms. Collectivization was forced, and two-thirds of the cooperatives disbanded spontaneously in 1956. The cooperative sector was subsequently expanded again, and by the late 1980s cooperative farms cultivated more than two-thirds of the land, while the share of state farms was more than one-fourth. Only 1.1 percent of the cultivated land was privately owned. Owing to mismanagement, however, one-third of the agricultural cooperatives were insolvent. The household plots of members of cooperative farms and auxiliary and individual holdings provided about one-third of Hungary's total agricultural output.

Cooperative and state farms

Industry. During the early 1950s industry concentrated on the production of capital goods, but from the late 1950s greater emphasis has been placed on consumer goods. The earlier emphasis on heavy industry, however, has meant that few Hungarian industrial products are competitive in Western markets. In the late 20th century industry underwent a major restructuring, eliminating unprofitable units and increasing the production of market-oriented products. The principal branches of industry are engineering and chemicals: the engineering sector employs about one-third of the industrial labour force and accounts for the largest proportion of output, and chemicals account for one-fifth of exports. Other significant industries are textiles, clothing, food processing, and paper products. Newly established industries include biotechnology, microelectronics, and telecommunications.

In the past Hungary depended very heavily on coal as a source of energy, but production began to decrease with the increased use of oil and natural gas. The lowland nature of the country and the absence of fast rivers make the generation of hydroelectricity difficult. Construction of the Gabcikovo (Bös)–Nagymaros hydroelectric power plant on the Danube along the border with Czechoslovakia may change this. Hungary's first nuclear power station, at Paks (on the Danube south of Budapest), generates almost two-fifths of the country's total electricity. Plans were announced in 1987 to add more plants, increasing nuclear energy's share of electricity production to 50 percent by the end of the 20th century.

Finance and trade. The financial system is centralized and mainly owned by the state. Reforms implemented in 1987 separated central and commercial banking. The Hungarian National Bank is concerned with money supply and oversees the banking system. It is also involved in the formulation of economic policy and the control of foreign exchange and foreign debt. The national bank's earlier control of credit operations was transferred to the commercial banks, which also accept accounts and deposits. Specialized institutions serve the needs of particular sectors of the economy. In the late 1980s there were also three "consortium" banks—that is, institutions partly owned by foreign financial interests. Government permission is necessary for the establishment of consortium banks. Savings banks and cooperatives provide banking services to the general public.

In the 1980s bonds (to finance housing and infrastructure development) and treasury bills (to finance the public debt) were issued. State guarantees for bonds were discontinued in 1988. In that year a national securities market opened in Budapest; a stock exchange was established there in 1989.

Before World War II, food products were the main exports. Postwar industrialization, however, meant a restructuring of foreign trade, and the import of capital equipment, raw materials, and semifinished products and the export of industrial articles became dominant. Hungary is an exporter of engineering products, machine tools, buses, telecommunications and electrical equipment, electronic instruments, pharmaceutical products, rolled steel, and clothing and footwear. Bauxite is a sizable export. Meat, fruits, vegetables, and wine are also significant, with agricultural products accounting for almost one-third of total exports. Principal imports are crude petroleum and natural gas (mainly supplied by pipeline from the Soviet Union), iron ore, copper, and consumer goods.

Foreign trade

The Soviet Union is by far the largest trading partner; among the capitalist countries, West Germany is the most important. Hungary's convertible currency trade position deteriorated with falling world market prices for agricultural exports and a decline in the forint's exchange rate against hard currencies. Regulations introduced in 1988 permitted all businesses to export their products and to conduct business with foreign partners without the involvement of government agencies.

Transportation. The industrialization and urbanization of Hungary have increased the importance of transport. The population has become more mobile while the demands of industry for the movement of raw materials

Milt and Joan Mann—Cameramann International



Harvesting corn near Dunaújváros, in central Hungary.

and finished articles have risen significantly. The growing tourist traffic between East and West also places greater burdens on the transport system.

Transportation in Hungary is strongly centralized, and both the main railway links and principal roads converge on Budapest. Roads have overtaken railways in terms of volume of freight transported; increasing numbers of people also use roads for personal transportation. In addition, transport by pipeline is developing.

Railways. The main elements of the Hungarian railway system were constructed in the second half of the 19th century to connect the principal provincial cities with Budapest, and, taking frontier changes into account, traffic is still concentrated in these few important lines. Domestically, the most important links are between Budapest and northern industrial areas; internationally, the main lines to the Soviet Union, Austria, Czechoslovakia, East Germany, and Poland carry the most traffic. Steam power has been virtually eliminated in favour of diesel traction, and the main east–west rail link from the Austrian to the Soviet frontier is electrified.

Modernization of railways

Roads. A majority of roads in the country have modern asphalt or concrete surfaces. The main roads radiate from Budapest, and there are extensive bus services. Car ownership is expanding dramatically; and the growth of domestic and international road haulage places severe pressure on parts of the system. The London–Damascus and Hamburg–Bucharest international highways cross the country. A four-lane expressway, from Budapest to Lake Balaton, is in operation.

Water transport. The Danube is the only waterway in the country transporting a significant amount of freight, much of it international. The principal port is the Csepel free port in Budapest. Tourism provides the bulk of a highly seasonal traffic on other waterways, Lake Balaton deserving special mention in this connection.

Air transport. There is no domestic air service within Hungary. International air traffic centred on Ferihegy Airport, Budapest, is rapidly expanding. Malév, the national airline, operates to almost every country of Europe as well as to the Middle East. Ferihegy is also used by a number of major Western airlines.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Until 1988 Hungary was a one-party state whose structure of government and laws were set out in the Soviet-type constitution of 1949 as revised in 1972. Both the legislative and executive branches of government were controlled by the Hungarian Socialist Workers' Party (HSWP; the only political party), which played the leading political and ideological role in the Hungarian People's Republic. Many members of its Political Committee were members of the government.

Creation of multiparty system

The draft of a new constitution approved by the National Assembly in the spring of 1988 provided for a multiparty system and the separation of powers, replacing the one-party state with a representative democracy. In October 1989 the Assembly enacted about 100 constitutional changes creating such a system. As part of the movement for reform, the HSWP changed its name to the Hungarian Socialist Party and became one of a number of parties contending for a position in Hungarian politics, including the Hungarian Democratic Forum, the Alliance of Free Democrats, and the Independent Smallholders' Party. A small group of conservative HSWP members retained the party's old name and continued to advocate its policies.

The supreme organ of state is the National Assembly (Országgyülés), which selects the president and the Council of Ministers, the president of the Supreme Court, and the chief public prosecutor. The premier, generally the head of the ruling party, is also selected by the National Assembly. In the past, when the National Assembly was not in session, its functions were executed by the Presidential Council. In 1989 the council was replaced by the creation of a state president, a relatively weak position. Real power lies with the premier and ruling party in the Assembly.

The main organ of state administration is the Council of Ministers, whose president and individual ministers have the authority to issue orders. The Council of Ministers delegates authority to commissions concerned with specialized spheres.

Local government. Local administration is based on a hierarchical structure of councils. At the base of the hierarchy are village councils, which are subordinate to the county council within whose geographic sphere they lie. Of equal rank to the latter are the municipal councils (subdivided into subordinate district councils) of Debrecen, Győr, Miskolc, Pécs, and Szeged. The administration of the capital is the responsibility of the Metropolitan Council of Budapest and the subordinate metropolitan district authorities. As a reform measure, councils of villages with small populations have been amalgamated into joint village councils, which makes possible the extension of services (such as health centres, libraries, and primary schools) to rural communities where previously there was no economic justification. The county and municipal authorities together with the Budapest metropolitan and metropolitan district councils represent the most important tier of local government. Their specialist executive committees deal with fiscal matters, economic administration (including the running of some industry and many retail functions), planning and control of education, and the supervision of public health.

County and municipal councils

Elections. Parliamentary and council elections are held every five years on a franchise extended to all Hungarian citizens over the age of 18. Members of the National Assembly and of village, municipal, urban, and metropolitan district councils are elected directly by secret ballot. Because the county and Budapest metropolitan councillors represent the interests of a whole city or county, they are elected by the members of the subordinate village and metropolitan district councils rather than by direct franchise. Candidates are chosen at nomination meetings, and more than one candidate can stand for election in a given constituency. When several candidates contest a constituency, the winner is the one who obtains an absolute majority of the votes cast. If no candidate succeeds in winning a majority in the first round of elections, a runoff of the parties gathering the largest number of votes is held.

Elections to the National Assembly are for seats allocated in three ways: most seats represent individual constituencies, somewhat fewer are decided on the basis of proportional representation, and the remainder are elected at large. Elected members of the National Assembly and of councils are accountable to their constituents, who have the right to remove them in the case of incapacity or lack of confidence. In such a case or when a member dies, a special election is held.

Justice. The administration of justice is the responsibility of the chief public prosecutor, who is elected for a six-year term by the National Assembly. Civil and criminal proceedings come under the jurisdiction of the lower courts and of the Supreme Court. The Supreme Court passes judgment on matters of principle, which are then binding on the other courts. Local government executive committees may try petty offenses, and the police have power to deal summarily with minor traffic infringements and other misdemeanours. Judges and lay assessors are elected by the local councils and members of the Supreme Court by the National Assembly, on the recommendation of the Presidential Council.

The highly politicized legal system failed to redress the lawlessness that victimized hundreds of thousands of citizens during the 1950s and early '60s. Although the revision of politically motivated or otherwise suspect court cases has begun, the constitutional court to be established under the new constitution will have the task of dealing with violations of citizens' rights. Indeed, in the effort to reestablish a rule-of-law state (*Rechtsstaat*), Hungary is in the process of reconsidering its entire legal and administrative system.

Armed forces. The armed forces consist of the People's Army, which includes the air force, the frontier guards, the police, and the customs police. All males over the age of 18 are liable for compulsory military training, the length of which varies according to the branch of the services entered. The police safeguard social property, public order, and the personal and material security of citizens

and are responsible for traffic control. The police force, which is also supported by voluntary groups, comes under the jurisdiction of the National Police Board.

Education. In Hungary schooling is free (financed out of general government funds, with parental contributions toward the cost of books and stationery) and compulsory between the ages of six and 16. Many children attend crèches and kindergartens before reaching the age of six. Elementary education extends over eight years and embraces the six-to-14 age group. At the age of 14, children embark upon their secondary education in grammar schools, technical schools, or schools specializing in the fine arts or music. Entrance into secondary school is based upon elementary school reports and parental choice, and most pupils stay on until the age of 18. Technical schools have become increasingly popular because, in addition to teaching traditional subjects, they also provide pupils with a vocational background. Each school is free to determine which subjects it considers compulsory and which elective. The teaching of Russian as a compulsory foreign language ceased in 1989.

Higher education

Institutions of higher education, of which there are more than 80, range from the traditional university to the very specialized college. Acceptance into a university is dependent upon the results of an entrance examination, although social background is also taken into account. Higher education is free, and some students are also provided with stipends. In the late 1980s, modernization of the Hungarian educational system slowed because of cutbacks in the educational budget, low teachers' salaries, and inadequate stipends at a time of inflationary pressure.

Health and welfare. Under the national health plan, medical treatment is available free of charge, as are drugs provided in hospitals, vaccine serums and lifesaving drugs, and drugs prescribed for minors. The health service is administered by local councils under the supervision of the Ministry of Social Welfare and Health. Great attention is paid to disease prevention through the screening of the population for tuberculosis and cancer and through mass immunization against smallpox, poliomyelitis, and tuberculosis. But medical care, previously of a high standard, has deteriorated. Doctors and nurses receive inadequate salaries, and gratuities are expected. Private practices are permitted and their number is growing, but not all patients can afford private care. Shortages in medical supplies and appliances are frequent.

Social insurance, which is paid for largely by the contributions of employers, provides for sickness, maternity, and death benefits. Sick pay is available for one year without restriction. In addition to 24 weeks' paid maternity leave, mothers can receive a monthly child-nursing grant for a period of two years after birth. Compensation for work injuries and unemployment compensation are also paid. Both employers and employees contribute to the pension fund, from which are drawn retirement (at 60 years of age for men and 55 for women) and disability pensions. The social and pension funds are administered by the National Social Security Council, under the supervision of the Council of Ministers. Retired persons who are not entitled to a pension, together with those whose pensions are too low, receive social care from the appropriate council. Councils also provide homes for the aged who have no immediate family and, in conjunction with employers, maintain nurseries and kindergartens as well as day-care centres for children of school age. Family allowances are paid on a progressive scale to married couples with two or more children and to single-parent families with one or more children. Children in need of care are looked after in homes maintained by the state. The rapid deterioration of the economy in the 1980s, however, was aggravated by budget cuts for social services, which have made the welfare and pension systems substandard.

The housing shortage

Housing. Housing remains a severe problem. The basic premise of housing policy is that every family is entitled to its own home at reasonable cost and each family member to a separate room. Overcrowding, shared tenancies, sublettings, and the involuntary sharing of accommodations with parents are nevertheless common. While the state is active in the housing field, the majority of dwellings are privately owned. Public-sector housing is controlled by local authorities. There is a multiplicity of home types. The state in the form of the local authority is the largest landlord, yet it builds homes for sale as well as for rent. The state and private sectors are bridged by the cooperative housing sector, although the cooperative scene also includes groups of individuals who voluntarily form themselves into a "private" cooperative to build a modest apartment house of not more than 10 flats. The privately owned single-family dwelling is most common in small towns and villages.

Deficiencies in housing policy regarding availability, upkeep, and social distribution of homes led to the financial involvement of individual households in home construction and maintenance. Rents skyrocketed, though they are still low by most Western standards. To reduce the burden on the state, private ownership is encouraged through the provision of subsidies and mortgages at low rates of interest. Housing policy and population policy are interconnected in that housing subsidies go to families with young children or to married couples who agree to have children within five years.

CULTURAL LIFE

The cultural milieu. A large number of Hungarians have made substantial contributions to the arts and sciences. The achievements of the Hungarian-born scientists Albert Szent-Györgyi, Georg von Békésy, Eugene Wigner, Edward Teller, Leo Szilard, and John von Neumann have gained world recognition and Nobel prizes. The music of Franz (Ferenc) Liszt, Béla Bartók, and Zoltán Kodály is also known throughout the world, and many contemporary Hungarian musical artists and conductors have become eminent internationally. In addition, the works of the poets János Arany, Sándor Petőfi, and Endre Ady as well as those of the writers Kálmán Mikszáth, Zsigmond Móricz, and Ferenc Molnár have been translated into many languages.

Before World War II, Hungary's well-developed cultural life was largely confined to the upper classes. The working class was regarded as uncultured, and illiteracy was widespread. Since 1945 the government has made great efforts to promote mass culture, and large sums have been and continue to be spent by the state in this respect. Cultural opportunities have been greatly expanded, and prices are kept low through state support of literature, music, the visual arts, and the dramatic and popular arts.

Mass culture

Hungary in general and Budapest in particular are noted for a vigorous cultural life. Hungarian performers often appear abroad and have achieved many successes in international competitions, while in the capital and other cities celebrated foreign artists often perform. Foreign exhibitions are also frequently put on for the benefit of the Hungarian public. The Hungarian cinema has received numerous awards at international film festivals, and films produced in some 20 countries are regularly shown to Hungarian audiences.

Amateur groups—numbering more than 10,000—are also active, and some preserve the Hungarian traditions of pottery making, carving, weaving, and embroidery. These crafts are also supported by the activities of art cooperatives, whose embroidery and lace are appreciated both inside and outside the country. Also well known are the handicrafts of Matyó-land, Sárköz, and Buzsák, the lace of Kalocsa, and the folk murals decorating peasant homes in Tolna and Bács-Kiskun counties. Other folk products include the pottery of Békés county and the glass of Nógrád.

Cultural institutions. Free public libraries are found in almost every town and village of Hungary, comprising more than 9,000 general and more than 1,500 specialized technical libraries. An important role in the spread of culture in the village is also played by the cultural centre and in the population at large by the more than 600 national museums. A dozen or so scientific societies exist, and the Hungarian Academy of Sciences maintains a number of research institutes. The country has many motion-picture houses. Traveling theatres perform in the villages, and the theatre companies of the large cities also go on tour. As in other countries, the spread of television has caused a

drop in theatre and cinema audiences and in attendance at cultural programs in general. This has generated a wide-ranging debate regarding the relative claims for state help of the different cultural forms.

A large range of daily and weekly newspapers and periodicals is published. Radio remains popular, but during the late 20th century it was displaced from its former prominent position by television. Although the press was subject to government control for more than 40 years, by 1988 Hungary had many independent papers and journals, and the media was free of censorship.

Sports

There are several thousand sports clubs in the country. Football (soccer) remains the most popular sport, but the country is also strong in swimming, rowing, boxing, pentathlon, wrestling, and weight lifting. Tennis has gained in popularity.

For statistical data on the land and people of Hungary, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL. (G.Ba./Ed.)

History

Hungary came into existence when the Magyars occupied the middle basin of the Danube in the late 9th century. Parts of its territory had formed the ancient Roman provinces of Pannonia and Dacia. When Rome lost control of Pannonia at the end of the 4th century, it was occupied first by Germanic tribes, then by Slavs. The subsequent history of Dacia is unrecorded. The central plains had formed the bases of nomadic immigrant peoples from the steppes north of the Black Sea—Huns, Bulgars, Avars—some of whom extended their domination farther afield. The Avars, who dominated the basin in the 7th and 8th centuries, were crushed in about 800 by Charlemagne, whose successors organized the western half of the area in a chain of Slavic vassal "dukedoms." One of these, Croatia, which extended as far north as the Sava, made itself fully independent in 869, while another, Moravia, which then extended as far as the Gran, had openly defied its Carolingian overlord for as long. The Byzantine Empire and Bulgaria exercised loose authority over the south and east of the area.

THE KINGDOM TO 1526

The history of the Magyars

The Árpáds. In 892 the Carolingian emperor Arnulf, attempting to assert his authority over the Moravian duke Sviatopluk, called in the help of the Magyars, a Finno-Ugric people, whose early homes had been on the upper waters of the Volga and Kama rivers; unrecorded causes had driven them, at an uncertain date, southward into the steppes, where they had adopted the life, appropriate to their environment, of peripatetic herdsmen. In the 9th century they were based on the lower Don, ranging over the steppes to the west of that river. They then comprised a federation of hordes, or tribes, each under a hereditary chieftain and each composed of a varying number of clans, the members of which owned a real or imagined blood kinship. The clansmen were all free, but the community included slaves taken in battle or in raids. There were seven Magyar hordes, but other elements were included in the federation, including three hordes of Turkic Khazars (the Kavars). Either from this fact or perhaps from a memory of earlier conditions, this federation was known to its neighbours as the On-Ogur (literally, "Ten Arrows"), from the Slavic pronunciation of which the name "Hungarian" is derived.

In 889, attacks by the newly arrived Pechenegs had driven the Magyars and their confederates to the western extremities of the steppes, where they were living when Arnulf's invitation arrived. The band sent to Arnulf reported back that the plains across the Carpathians would form a suitable new homeland that could be easily conquered and defended from the rear. Having elected as their chief Árpád, the leader of their most powerful tribe, the Magyars crossed the Carpathians en masse, probably in 896, and easily subjugated the peoples of the sparsely inhabited central plain, their first place of settlement. They destroyed the Moravian empire in 906 and in the next year occupied Pannonia, having defeated a German force sent against them. They were then firmly established in the whole centre of the basin, over which their tribes and their associates distributed themselves, Árpád taking the central area west of the Danube for his own tribe. The periphery was guarded by outposts, which were gradually pushed forward, chiefly to the north.

The Christian kingdom. During the next half century the Magyars were chiefly known to Europe for the forays they made across it, either as mercenaries in the service of warring princes or in search of booty for themselves—treasure or slaves for domestic use or sale. But terrifying to others, their mode of life was not always profitable to themselves. Their raiding forces suffered several severe reverses, culminating in a disastrous defeat at the hands of the emperor Otto I in 955 on the Lechfeld, outside Augsburg. By this time the wild blood of the first invaders was thinning out and new influences, in particular that of Christianity, were beginning to operate. Both the Eastern and Western churches strove to draw them, with the other peoples of east central Europe, into their orbits. They had established pacific, almost friendly relations with Bavaria. The decisive step was taken by Árpád's great-grandson, Géza, who succeeded to the hereditary chief leadership in 972 and reestablished its authority over the tribal chiefs. In 973 he sent an embassy to Otto II at Quedlinburg, and in 975 he and his family were received into the Western church. In 996 his son, Stephen (István), married Gisella, a Bavarian princess.

Magyar defeat at Lechfeld

Stephen I (997–1038) carried on his father's work. With the help of heavily armed Bavarian knights, he crushed his rivals for the headship. Applying to the Holy See, he received the insignia of royalty from it and, according to tradition, was crowned king on Christmas Day, 1000. The event was of immeasurable importance, for not only did Hungary now enter the spiritual community of the Western world but Pope Sylvester's personal relationship with his former pupil, Otto III, enabled it to do so without admitting the political suzerainty of the empire. Stephen then carried through the conversion of his people to Christianity, establishing a network of archiepiscopal and episcopal sees, which he reinforced with lavishly endowed monastic foundations. Later he crushed the surviving disputants of his authority—notably the Kavars—and, carrying on his father's work, organized his state on a system that was to remain for many centuries the essentially unaltered basis of Hungary's political and social structure. The tribes, as units, disappeared, but the fundamental social stratification was not altered. The descendants in the male line of the old conquerors, and elements later equated with them, remained a privileged class, personally free, answerable in judgment only to the king or his representative, and entitled to appear in general assemblage. Their lands—which at this time, since the economy was mainly pastoral, were held by clans or subclans in semicommunal ownership—were inalienable, except for proved delinquency, and free of any obligation. The only duty required by the state of members of this class was that of military service on call. They were allowed to retain their slaves, although Stephen freed his own. All land outside that held by this class—then more than half the whole—belonged to the crown, which could, indeed, donate it at will. The nonservile inhabitants of these lands—descendants of the pre-Magyar population, manumitted slaves, invited colonists, etc.—were subjects of the crown or of the local landholder.

Conversion to Christianity

The whole of this land was divided into counties (*megyék*), each under a royal official, an *ispán* (later *fő* [head] *ispán*) who represented the king's authority in it, administered its unfree population, and collected the taxes that formed the national revenue, central and local. Each *ispán* maintained at his headquarters, or *vár* (fortress), an armed force composed of freemen who took service under him or of persons freed by the king. In Stephen's day there were about 45 such counties.

The early kings. Once St. Stephen (he was canonized in 1083) established his rule, his authority was rarely questioned. He fought few foreign wars and made his long reign a period of peaceful consolidation. But his death in 1038 was followed by many years of discord, for his only son, Imre (Emeric), had predeceased him, and the nation

A period of disputed succession

rebelled against his designated successor, Peter, who was expelled in 1041. Peter returned in 1044, with the help of Emperor Henry III. The "national" king, Samuel Aba, who had taken his place, was murdered, but Peter himself was killed in another rebellion in 1046. Andrew (Endre) I, of a collateral branch of the house of Árpád, was killed in 1060 fleeing from his brother, Béla I, and after Béla's death there was a further conflict between his sons, Géza and Ladislas (László), and Andrew's son, Salamon. Peace returned only when, after Géza's short rule (1074–77), the throne passed to Ladislas I, who occupied it until 1095; and even then, the curse of dynastic jealousy proved to have been exorcised only temporarily. Ladislas' successor, Coloman (Kálmán; 1095–1116), who was Géza I's elder son, had his own brother, Álmos, and Álmos' infant son, Béla, blinded to secure the throne for his own son Stephen II (1116–31), perhaps Hungary's worst king. Béla II (1131–41; the blinded boy, whom his father's friends had brought up in secrecy) and Béla's eldest son, Géza II (1141–61), ruled thereafter unchallenged, but the succession of Géza's son, Stephen III (1161–73), was disputed by two of his uncles, Ladislas II (1162–63) and Stephen IV (1163–65). Happily, the death of Stephen IV exhausted the supply of uncles, and Stephen III's brother, Béla III (1173–96), had no domestic rivals to the throne, but the short reign of Béla's elder son, Emeric (Imre; 1196–1204), was spent largely in disputes with his younger brother, Andrew II, who on Emeric's death expelled his infant son, Ladislas III (who died the next year), before beginning his own long reign (1205–35).

Consolidation and expansion. These royal disputes caused Hungary much harm. Claimants to the throne often invoked foreign help, for which they paid in political degradation or temporary or permanent loss of territory: both Peter and Salamon did homage to the Holy Roman emperor for their thrones; and Aba's war against Peter's protectors cost Hungary its previous territories west of the Leitha, while the wars of the 12th century cost it areas in the south. The uncertainty delayed political consolidation, and even Christianity did not take root easily: there was a widespread pagan revolt in 1047 and another in 1063.

Yet the political unity of the country and the new faith somehow survived the earlier troubles, and both were firmly established by Ladislas I, one of Hungary's greatest kings, and by Coloman, who, in spite of his atrocious crime, was a competent and enlightened ruler.

Meanwhile, many factors worked for Hungary. After Austria had grown big at the expense of the imperial authority, most of Hungary's neighbours were states of approximately the same size and strength as itself, and the Hungarians lived with them on terms of mutual tolerance and even friendship. The steppes were quiet: the Cumans (Hungarian: Kun), after destroying the Pechenegs there, did not try to go farther, and, after two big raids had been successfully repelled by Ladislas, they left Hungary in peace. This allowed Hungary to extend its effective frontiers to the Carpathian crest in the north and over Transylvania. Magyar advance guards pushed up the valleys of both areas, and these were reinforced in the Szepes area and central Transylvania by imported colonies of Germans (usually called "Saxons"), while colonies of Szeklers (Székely), a people akin to the Magyars who had preceded the latter into the central plains, were settled behind its eastern passes. The county system was extended to both areas, although with modifications in Transylvania, where the Saxons and Szeklers constituted free communities and the whole was placed under a vaivode, or governor. In the south Ladislas occupied (or reoccupied after an interval) the area between the Sava and Drava rivers; Coloman assumed the crown of Croatia, which then included Bosnia and Northern Dalmatia, although this remained a separate "Land of the Hungarian Crown," over which a governor (ban) acted as deputy for the king.

Imported colonists

In the interior, too, natural growth and continued immigration swelled the population, which by 1200 had risen to the large figure, for the age, of some two million. The rulers of this big and populous state were now important men. After Ladislas' day, nothing more was heard of German claims to suzerainty over it. In the 12th century Hungary was intervening in its neighbours' affairs as often as they in Hungary's. Béla III, who married a French princess, had the largest revenues of any European monarch, after the two emperors.

Social and political developments. Meanwhile, the pattern of Hungarian society had been changing. The numbers of the free class, or "nobles" as they were coming to be called, although frequently reinforced by new admissions to its ranks, probably hardly increased in absolute terms, and certainly far less than those of the unfree population: from perhaps half the total in 896, they had been reduced by 1200 to about one-eighth. Further, as the economy became agricultural, the old clan lands dwindled until only pockets remained. Where the rest had been and in large parts of the old crown lands, which improvident donations had greatly reduced, the land was held in the form of individual estates, the owner of each of which was master of the unfree population on it: the nobles had, to a large extent, become a landed oligarchy. Some individual estates were very large, and their owners had come to constitute a "magnate" class, not yet institutionalized or legally differentiated from their poorer co-nobles but far above them in wealth and influence. The non-nobles, although slavery had practically disappeared, were still a "subject" class; and although many of them, including the burghers of the towns (most of which were German foundations) and such communities as the Saxons and Szeklers were protected by special charters and personally free, even they stood, politically, outside the magic ring of the *natio Hungarica*.

Throughout these developments the political form of the country had remained that of the absolutist patrimonial kingship. The king maintained a council of *optimates* (aristocrats), but his prerogatives were not restricted and his authority on matters falling within it remained absolute. A strong king, such as Béla III, could always curb a recalcitrant magnate by simply confiscating his estate. Only the follies and extravagances of the feckless Andrew II evoked a revolt, culminating in 1222 in the issue of the Golden Bull, to which every Hungarian king thereafter had to swear. Its purpose was twofold: to reaffirm the rights of the smaller nobles of the old and new (*servientes regis*) classes against both the crown and the magnates and to defend those of the whole nation against the crown by restricting the powers of the latter in certain fields and legalizing refusal to obey its unlawful commands (the *jus resistendi*). But Andrew had done much harm by dissipating the royal revenues through his extravagances and by huge grants of land to his partisans.

Issuance of the Golden Bull

The Mongol invasion: the last Árpád kings. Andrew's successor, Béla IV (1235–70), began his reign with a series of measures designed to reestablish royal authority, but his work was soon interrupted by the frightful disaster of the Mongol invasion. In the spring of 1241 the Mongols quickly overran the country and before they left it, a year later, had inflicted ghastly devastation. Only a few fortified places and the impenetrable swamps and forests escaped their ravages. The country lost about half its population, the incidence ranging from 60 percent in the Alfold (100 percent in parts of it) to 20 percent in Transdanubia; only parts of Transylvania and the northwest came off fairly lightly. Returned from Dalmatia, where he had taken refuge, Béla, whom his country not unjustly dubbed its second founder, reorganized the army, built a chain of fortresses, and called in new settlers to repopulate the country. He paid special attention to the towns. But he was forced to give some of the magnates practically a free hand on their own estates, and a few families rose to near-sovereign local status. Further, one group of immigrants, a body of Cumans who had fled into Hungary before the Mongols, proved so powerful and so turbulent that to ensure their loyalty Béla had to marry his son, Stephen V, to a Cuman princess. Stephen survived his father only two years, after which the country passed under the regency of his widow, the "Cumanian woman," whom the Hungarians detested. Her son grew up into a wild, undisciplined youth, who took pleasure only in Cuman mistresses and who came so near to paganism that Pope Nicholas IV preached a crusade against him. He was assassinated (by

Rebuilding under Béla IV

a Cuman) and left no legitimate heir. Such of Hungary's neighbours as could pretend to the succession by descent from an Árpád through the female line prepared to put in their claims. One male of the Árpád line was, however, discovered in Italy. His legitimacy was impugned, but his supporters got him into the country, and the young man, Andrew III, proved a wise and capable king. Unhappily, he died in 1301, without male issue, and with him the national dynasty became extinct.

Hungary under foreign kings. The extinction of the old dynasty created a new situation for Hungary. The nation was now entitled to choose its successor; but the principle of the blood tie was still generally regarded as determinant, and all the candidates for the throne—Wenceslas of Bohemia, Otto of Bavaria, and Charles Robert of the Angevin house of Naples—based their claims on descent from an Árpád in the female line. But all three claimants were foreigners; one of them and the father of another were actually seated on foreign thrones. From that time until its extinction, the kingship of Hungary was in fact invariably—with two exceptions, one of them disputed—held by a foreigner, nearly always by one occupying simultaneously at least one foreign throne. This could be to the advantage of Hungary if he used the resources of those thrones in its service, but he could, alternatively, neglect and exploit Hungary in his other interests and use his power to crush its national freedoms and institutions. Securing the advantages of foreign rule while escaping its dangers was the abiding dilemma—seldom successfully resolved—of Hungarian history.

The Angevin kings. The problem did not pose itself at first, for Charles Robert of Anjou (1308–42), still a child when his supporters won out, had no foreign throne and grew up a true Hungarian. He was also a capable man. After growing to manhood, he crushed the most rebellious of the "kinglets" and won over the rest; after this his rule was unquestioned and peaceful at home. The international situation, with Germany distraught by the power struggle between empire and papacy, the Mongolian Tatars grown passive, and the power of Byzantium in full decay, was again favourable to the states of the "middle zone" of eastern Europe and the Balkans; it is no accident that Poland, Hungary, Bohemia, and Serbia all look on the 14th century as their golden age. As this situation favoured Hungary's neighbours, as well as itself, Charles Robert's attempts at expansion were only moderately successful. In the Balkans he made Bosnia his friend and client but lost Dalmatia to Venice and other territories to Serbia and the newly emerged *vaivody* (province) of Walachia. But he drove Czech and Austrian marauders out of the land and on the whole preserved friendly relations with Austria, Bohemia, and Poland.

Charles's son, Louis (Lajos) I (1342–82), the only Hungarian king on whom his country has bestowed the name "Great," built on his father's foundations. Keeping peace with the West, he repaired his father's losses in the south and surrounded his kingdom with a ring of dependencies

Adapted from R. Treharne and H. Fullard (eds.), *Muir's Historical Atlas: Ancient, Medieval and Modern*, 9th ed. (1962); George Philip & Son Ltd., London



Hungary in 1360.

over which Hungary presided as *archiregnum* (chief kingdom) in the Balkans, on the lower Danube, and in Galicia. In 1370 he also ascended the throne of Poland, in virtue of an earlier family compact.

Angevin prosperity

Both Angevin kings owed much to the wealth they derived from the gold mines of Transylvania and northern Hungary, some 35 to 40 percent of which went to the king, enabling him to maintain a splendid court. Spared for two generations from serious invasion or civil war, the rest of the country blossomed materially as never before. The population rose to three million and the country contained 49 royal free boroughs, more than 500 smaller towns, and some 26,000 villages. The economy was still mainly rural, but the crafts prospered, trade expanded, and the arts flourished; a university, one of the oldest in Europe, was founded at Pécs in 1367 (though it proved to be short-lived).

The rule of the two Angevin kings was essentially despotic, although enlightened. Themselves no bureaucrats, they introduced elements of feudalism into the political, and especially the military, system: each lord was responsible for maintaining his own armed contingent (*banderium*). But the magnates were held firmly in check, and Louis reaffirmed the rights and privileges of the common nobles and carried further a process, begun in the previous century, under which the counties were developing from "royal" into "noble" institutions, each still under a royal official but administered with a wide measure of autonomy by elected representatives of the local nobility. He also standardized the obligations of the peasants at the figure of one-ninth of their produce to the lord, a tithe to the church, and a house tax (*porta*) that went directly to the state.

Sigismund of Luxembourg. The benefits of Louis's rule would have been far greater still had he not wasted enormous sums of money and many lives on endeavours to secure the throne of Naples for his nephew. His foreign acquisitions served his personal glory more than they did the real interests of his country; much of the imposing edifice collapsed when he died, leaving as issue only two daughters. Louis had destined the elder, Maria, to succeed him on both his thrones, but the Poles refused to continue the union. They accepted the younger daughter, Jadwiga (Hedvig), as queen but married her to Jagiełło of Lithuania. The Hungarians crowned Maria, whose husband, Sigismund of Luxembourg, had himself crowned as her consort in 1387 and after her death eight years later ruled alone until his own death in 1437. Under him, matters took a sharp turn for the worse. Sigismund did much for arts, commerce, and, above all, for the towns. Also, like Andrew II, he promoted Hungarian political institutions by creating the need for them. The principle that the consent of representatives of the privileged classes, assembled in the Diet, was necessary for the grant of any subsidy or additional taxation, and even, later, for any legislation, dates from his reign, being made necessary by his extravagance and arbitrariness. His frequent and prolonged absences called into being the peculiar Hungarian institution of the palatine, an officer elected jointly by the king and the nation, who represented the king during his absences and also acted as intermediary between him and the nation. But these were only palliatives against abuses that were bitterly felt. The nation, besides hating him for the cruelty he showed at the outset of his reign against the supporters of a rival, resented bitterly the absenteeism of his later years, when he occupied himself chiefly with imperial and Bohemian affairs (he was elected German king in 1411 and became titular king of Bohemia in 1419), neglecting—it felt—the numerous problems of Hungary. There was much discontent among the peasants, who were subjected to heavy exactions both by the crown and by their masters, the unrest being aggravated by the spread of radical Hussite religious doctrines from Bohemia; there were serious revolts in northern Hungary and Transylvania. Above all, there was the growing danger from the Ottoman Turks, who, though they had already taken Bosnia from Louis, could not threaten Hungary proper while Serbia still stood. But in 1389 the power of Serbia was broken at the Battle of Kosovo, and the danger became urgent.

The growing Ottoman threat

Sigismund organized a crusade, but this was disastrously defeated at Nicopolis in 1396. Timur (Tamerlane) gave Europe a respite by his attack on the Turkish rear, but the advance was resumed in 1415. Walachia submitted in 1417; thereafter Transylvania and southern Hungary suffered repeated raids.

János Hunyadi and Matthias Corvinus. The Ottoman sultan Murad II was preparing a grand assault on Hungary when Sigismund died in 1437, leaving as issue a daughter, married to Albert V of Austria. The country accepted Albert as Sigismund's successor, but only on condition that he should not become Holy Roman emperor and should not reside outside the country without permission of the estates. Albert set about organizing the country's defenses but, disastrously, died in 1439, leaving his widow and unborn child. To avoid an interregnum and a minority, perhaps with a queen, the country elected Władysław III of Poland as king; within two years of his death in battle against the Turks in 1444, they accepted Albert's son, Ladislas V (called Ladislas Posthumus), appointing as his guardian and governor of Hungary the great general János Hunyadi, who had been successfully repelling the renewed Turkish attacks. Hunyadi kept up the defense under increasing difficulties, constantly thwarted by jealous magnates and harassed by the Czech condottiere Jan Jiskra, while the emperor Frederick III encroached on the western provinces.

Hunyadi died in 1456 after the recapture of Belgrade from the Turks. Ladislas' maternal uncle, Ulrich of Cilli, aware of the country's devotion to Hunyadi, had the latter's elder son assassinated and his younger son Matthias (Mátyás) Corvinus imprisoned in Prague. Ladislas V himself died suddenly, still unmarried, a year later. The country was tired of foreign rule and its agents, and on Jan. 24, 1458, a great concourse of nobles acclaimed Matthias king. Extracted from Prague with some difficulty, he was brought to Buda and crowned amid nationwide rejoicing.

The only national king to reign over all Hungary after the Árpáds, Matthias Corvinus has been seen through something of a golden haze by historians. A true Renaissance prince, he was a fine natural soldier, a first-class administrator, an outstanding linguist, a learned astrologer, and an enlightened patron of the arts and learning. His collections of illuminated manuscripts, pictures, statues, and jewels were famous throughout Europe. Artists and scholars were welcomed at his court, which could vie in magnificence with any on the continent. Sumptuous buildings sprang up in his capital and other centres.

Matthias' despotism

Politically, too, he represented the ideas of the Renaissance. He listened to his council, convoked the Diet regularly, and actually enlarged the autonomous powers of the counties. But at heart he was a despot; his real instruments of government were his secretaries, men picked by himself, usually young and often of humble origin. His rule was in the main an efficient and, on balance, a benevolent one. He simplified and improved the administration and, above all, the laws, enforcing justice with an even hand. The debit side of his rule was the increased taxation imposed by him for his administrative innovations, his collections (which cost his subjects vast sums), and above all, the mercenary standing army, 30,000 strong (largely composed of the defeated Hussites and known after its commander, "Black John" Haugwitz, as the Black Army), which he kept as part of the royal *banderium* for use against enemies, at home and abroad.

At first he had much need for such a force, for, although the Turks were quiescent for a decade, there were discontented magnates, and both the Czechs and the Austrians were unquiet neighbours. But, after Matthias had crushed, expelled, or bought off these enemies, had built a chain of fortresses along the southern frontier, and had even reestablished a nominal and, in practice, worthless suzerainty over Bosnia, Serbia, Walachia, and Moldavia, he let himself be drawn into an ever-widening circle of campaigns against Bohemia and Austria. In 1469 he made himself master of Moravia, Silesia, and Lusatia, with the title (although this was also borne simultaneously by Podiebrad) of king of Bohemia, and in 1478 he forced Frederick to cede him Lower Austria and Styria. He argued to his subjects that his neighbours were untrustworthy and that he could not organize the great crusade against the Turks without the resources of the imperial and Bohemian crowns. But they were unconvinced, and in 1470 a party actually conspired to depose him in favour of a Polish prince. This enterprise collapsed, and Matthias next entered on a complex transaction with the new emperor, Maximilian I, under which his illegitimate son John (he had no legitimate issue) was to marry Maximilian's daughter in return for re-cession of the Austrian provinces and Maximilian's recognition of John. But on May 6, 1490, when on his way to the meeting that should have sealed the bargain, Matthias died suddenly, and the whole enterprise collapsed.

The Jagiellon kings: national decay. The magnates, who did not want another heavy-handed king, procured the accession of Vladislas II, king of Bohemia (Ulászló II in Hungarian history), precisely because of his notorious weakness—he was known as King "Dobre" (meaning "Good" or, loosely, "OK") from his habit of accepting with that word every paper laid before him. The emperor Maximilian contented himself with reoccupying his lost provinces and establishing a sort of paternal patronage over Hungary. This was consolidated in 1515 by an agreement under which Vladislas' son, Louis, married Maximilian's granddaughter, Mary, while Louis's sister, Anne, married Maximilian's grandson, Ferdinand, who was to succeed to Louis's thrones if Louis died without issue. The agreement was made over the heads of the Hungarians, and the Diet in 1505 passed a resolution never again to accept a foreign king. The candidate of the "national party" was János Zápolya, vaivode of Transylvania.

Meanwhile the magnates had disbanded the Black Army (without replacing it) and allowed the country's fortresses to fall into disrepair. Vladislas was their helpless prisoner; he could make no decision without their consent, and his revenues were looted so ruthlessly that he was reduced to selling Matthias' collections. Nearly all of Matthias' reforms were cancelled, and the peasants were oppressed grievously. In 1514 a great mass of peasants, led by a Szekler soldier, György Dózsa, rose against their lords. The revolt was put down with ghastly barbarity. The Diet of 1514 sentenced the peasants to "real and perpetual servitude" and bound them irrevocably to the soil (though this sentence was repealed only a few years later) and increased their dues and obligations.

When Vladislas died in 1516, his nine-year-old son was proclaimed king as Louis II. The defenses of the kingdom worsened, and in 1521 the new Ottoman sultan, Süleyman the Magnificent, sent Louis a demand for tribute; when this was rejected, Süleyman took Belgrade. Suddenly alive to the Turkish danger, the magnates voted the reestablishment of a standing army, but nothing was done to raise it, since each rival faction tried to put the burden of its upkeep on the others. Appeals for help from abroad met with little response. In 1526 the sultan advanced into Hungary. The general call to arms was proclaimed, but the most important forces—those from Transylvania and Croatia—were late in obeying it. Louis, with a force of 16,000 men, moved down the Danube, made contact with the Turks at Mohács on August 28, and attacked the next day. The Hungarian army, heavily outnumbered, was almost annihilated. Louis himself was drowned in his flight. Unable to believe that the pitiful array that had met him was Hungary's national army, the sultan advanced with extreme caution. He occupied Buda on September 10 but returned across the Danube by the end of October, taking with him more than 100,000 captives.

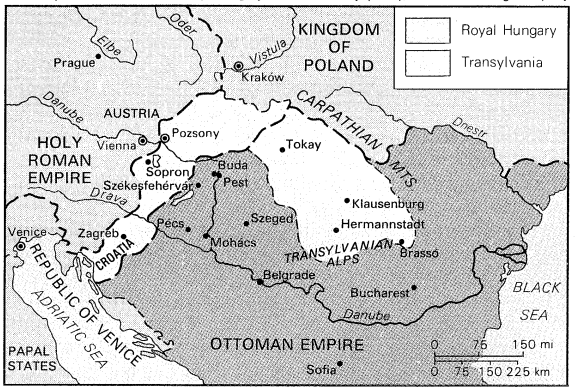
Renewed Ottoman threat

THE PERIOD OF PARTITION

Since the sultan had not meant to remain in Hungary, the disaster of Mohács would not have been irreparable had the king not perished; but, as it was, both Zápolya and Ferdinand of Habsburg (later Holy Roman emperor as Ferdinand I) claimed the throne, and each had supporters in the country. After each had failed to drive his rival out, Zápolya appealed to the sultan, who installed him in Buda, confining Ferdinand's rule to the western third of the country. A secret agreement was mediated in 1538 by Zápolya's adviser, György Martinuzzi ("Friar

George"), that Ferdinand should succeed to the whole heritage when Zápolya, the older man, died. The agreement was upset when, just before Zápolya died, his wife bore a son whom the national party recognized as king. The sultan then decided to act for himself. He recognized the infant as king but as his own vassal and further occupied Buda himself and incorporated a great wedge of central and southern Hungary in his own dominions. Ferdinand had to conclude a truce and to pay a tribute in return for recognition of his de facto rule over the territory then held by him. The transaction was completed when in 1566 the sultan formally declared Transylvania an autonomous principality under his own suzerainty; two years later Ferdinand's successor, Maximilian II, was forced to recognize this arrangement and to accept the reduction of Royal Hungary (that part of Hungary under Habsburg rule) to the western fringe of the country, the northwestern mountains, and Croatia.

Adapted from C.A. Macartney, *Hungary: A Short History* (1962); Aldine Publishing Company



The partition of Hungary in 1568.

The "age of trisection" was the blackest in all Hungarian history. The fighting and slave raiding, which went on even in times of nominal peace, reduced the whole south of the country to a wasteland occupied by only a few semi-nomadic Vlach herdsmen; villages disappeared and fields reverted to swamp and forest. Behind the new frontier the population was partially preserved to supply the garrisons, but the old landholders were replaced by Turkish officials and soldiers who, since their fiefs were not hereditable nor even always long-term, exploited the wretched cultivators to the maximum. Conditions were relatively tolerable only in those *kazalar* (districts) managed directly by the Ottoman government. In these districts, most of which lay along the two banks of the Tisza, the people flocked into the great "village towns" that are still a feature of the area. There they enjoyed a measure of protection; but the country between these towns was left empty except for scattered huts (*tanyas*) in which the men spent the summer scratching a precarious living from the soil.

Religious toleration

The Turks left Transylvania relatively unmolested. Martinuzzi devised a constitution based on earlier institutions, consisting, under the prince, of representatives of the three "nations": the Hungarian nobles of the counties, the Saxons, and the Szeklers. Transylvania was also spared internecine religious strife, when the representatives of the Roman Catholic, Calvinist, Lutheran, and Unitarian churches agreed on a regime under which these four confessions ranked as "established" on a basis of equal freedom and mutual toleration. The Greek Orthodox faith of the Romanians, who constituted the rest of the population, was only "tolerated," since the Romanians as such, or even their nobles, did not constitute a "nation."

Royal Hungary and the rise of Transylvania. In the first years after his accession, Ferdinand still hoped to bring the whole kingdom under his rule. He respected its constitution and its institutions and convoked the Diet regularly. But his hopes faded, and, after his succession to the imperial crown in 1558, Royal Hungary became no more than a small outlying annex of his mighty dominions. As it was also an exposed one, without the resources to defend itself, Ferdinand and his successor, Maximilian II, organized a chain of fortresses, mostly garrisoned by German troops, and a defensive "military frontier," inhabited by Serb and Vlach refugees from the Balkans and administered from Vienna. The Hungarians complained that they were being ruled and exploited as a subject people by foreigners, while Vienna looked on them as truculent rebels. Matters grew worse when Maximilian was succeeded by the unbalanced Rudolf, whose advisers hated Hungary and its traditions; and a religious conflict supervened on the constitutional dispute, for in the preceding half century the Reformation had swept over Hungary.

Religious antagonism played an important part when war between the empire and the Turks broke out again in 1591. In the so-called Fifteen Years' War imperial troops entered Transylvania and their commander, George Basta, behaved there (and in northern Hungary) with such insane cruelty toward the Hungarian Protestants that one of the Transylvanian generals, István Bocskay, formerly a Habsburg supporter, revolted. He improvised an army out of the wild herdsmen (*hajduks*), drove out Basta, and on June 23, 1606, concluded with Rudolf the Peace of Vienna, which left him prince of an enlarged Transylvania and also guaranteed the rights of the Protestants of Royal Hungary. He then mediated the Peace of Zsitvatorok (Nov. 11, 1606) between the emperor and the sultan, which kept the territorial status quo but relieved the emperor of his tribute to the sultan.

These two treaties ushered in a new era. The power of the Turks had begun to wane, and their place in the power equation was taken by Transylvania, which entered a half century of prosperity. Bocskay died suddenly in 1606, and the usual scramble for power followed; but in 1613 the Sublime Porte (the Ottoman government) imposed the election of the man who proved the most famous of all the princes of Transylvania, Gábor Bethlen (1613–29). At home Bethlen's rule was thoroughly despotic: through his monopoly of foreign trade and his development of the principality's internal resources, he almost doubled his revenues, devoting the proceeds partly to the upkeep of a sumptuous court, partly to the maintenance of a standing army. Keeping peace with the Porte, he often intervened against the emperor in the Thirty Years' War and safeguarded the rights of the Protestants in Royal Hungary. Under the Treaty of Nikolsburg (Dec. 31, 1621) he retained the title of prince of Transylvania and Hungary (the Porte vetoed his acceptance of the Hungarian crown), and a big extension of the principality, and a duchy in Silesia, besides further guarantees for the Protestants of Royal Hungary. When Bethlen died suddenly in 1629, his subjects abolished most of his internal reforms, but his successor, György Rákóczi I, maintained the international position of Transylvania, which figured as a sovereign state in the Peace of Westphalia (1648), ending the Thirty Years' War. This support from Transylvania and also the divisions prevailing between their own members prevented the Habsburgs from enforcing the Counter-Reformation in Hungary as early and as fully as they did in Austria and Bohemia. Nevertheless, the genius of the cardinal-primate Péter Pázmány won over for Roman Catholicism the majority of the local magnates, who came to form a party attached to the Habsburg cause, which was the more influential because they now formed a separate "table" of the Diet. The nation was thus divided, not only vertically (Transylvania versus the West) but also horizontally, between the Roman Catholic magnates and their subjects on the one hand and the smaller landowners, most of whom remained Protestants, on the other. In religious matters, the Hungarian Roman Catholic nobles were no more tolerant toward their Protestant fellow countrymen than were the emperor's own German and Czech advisers, although they were not willing to acquiesce in the political centralization championed by the latter.

Reign of Gábor Bethlen

War and liberation. The running sore remained the Turkish occupation of central Hungary, for every Hungarian resented the policy of the dynasty of leaving the Turks unmolested while pursuing ambitious policies in the west. It erupted in 1657 when György Rákóczi II of Transylvania, who had succeeded his father in 1648, allowed the prospect of obtaining the crown of Poland to seduce him into sending across the Carpathians an expeditionary

Reassertion of Turkish power

force, which was annihilated by Tatars. The grand vizier Mehmed Köprülü, the architect of the Porte's renaissance, led a force against Transylvania, detached it from the western adjuncts that had been its strength, and installed a new puppet prince. Emperor Leopold sent a force against the Turks; but, although the Austrian general Raimondo Montecuccoli defeated the Turks at Szentgotthárd on Aug. 1, 1664, the subsequent Peace of Vasvár recognized all the sultan's gains.

Now even the highest magnates of Royal Hungary plotted to expel the Habsburgs with Turkish and French help, but the conspiracy was betrayed and Vienna took its revenge. Nobles were executed or lost their estates, and Protestant pastors were sentenced to be galley slaves. In 1673 the constitution was suspended and Hungary placed under a directorate. A young Transylvanian, Imre Thököly, led a revolt that forced Leopold in 1681 to restore the constitution and revoke many of his harshest measures. The Porte, encouraged by Thököly's successes against the empire, sent into Hungary a vast army that in 1683 reached the walls of Vienna itself. But the tide ebbed as swiftly as it had advanced. Vienna was relieved, the Turks routed, and the imperial general Prince Eugene of Savoy led a series of campaigns in which all western and central Hungary, with Buda, were cleared of the Turks in 1686. Transylvania was liberated in the years following. By the Treaty of Carlowitz (Sremski Karlovci; Jan. 26, 1699), the sultan relinquished all of Hungary except the corner between the Maros and the Tisza rivers. (This was ceded in 1718 but kept until 1779 under Austrian administration as the Bánát of Temesvár.) The military frontier (progressively extended) was kept under a similar regime, and Transylvania was organized as a separate "principality."

HABSBURG RULE, 1699–1918

Habsburg rule to 1867. The emperor, not Hungary, was the victor, for the retreating Turks and the advancing armies of the so-called liberators ravaged the country, and, although in 1687 Leopold reconfirmed the constitution subject to Hungary's acceptance of his dynasty in the male line and to the abolition of the *jus resistendi* conceded under Andrew's Golden Bull of 1222, the government that followed was in fact another cruel centralist dictatorship. In 1703 this provoked another rebellion, led by Ferenc Rákóczi II; it took eight years of indecisive and fruitless fighting until peace was established by the Treaty of Szatmár (April 30, 1711). On paper, this did little more than confirm what had been agreed in 1687, but the new king Charles III (Emperor Charles VI) genuinely wanted peace with Hungary, and the worst abuses were now ended.

Charles III and Maria Theresa. Charles's chief concern was to secure the acceptance in Hungary of the Pragmatic Sanction, the imperial decree by which his daughter Maria Theresa should inherit his dominions. After the Diet accepted the Pragmatic Sanction in 1723, Charles convoked it only once again, and Maria Theresa, after her coronation in 1740, only twice more—each time to ask for money. Her rule, like her father's, was essentially autocratic. She was severe toward the Protestants, and she allowed her advisers to exclude Hungary from the subsidized industrialization that was bringing wealth to other parts of her dominions. But her rule was not more than severe, even toward the Protestants. Toward the magnates, on whom she lavished many favours, it was positively benign, and she respected the most cherished liberty of the lesser nobles, their exemption from taxation. Exhausted by so many wars and rebellions, the country asked for nothing more, contenting itself with the supreme blessing that her rule brought it an uninterrupted peace that enabled the population to grow once again and the material ravages to be repaired. But a lethargy descended on the country. Political life sank to the parish-pump level, and the towns stagnated. The peasants, into whose conditions the queen introduced some improvements, followed their masters in aspiring to nothing more than as much material comfort as could be obtained with a minimum of effort. The national language itself was becoming little more than a peasant dialect, since the language of public administration and the Diet was Latin and of business life, German; and with the language, the national spirit seemed near moribund.

The Pragmatic Sanction

Joseph II and Leopold II. The nation was shocked out of its lethargy by the accession of Maria Theresa's son, Joseph, on her death in 1780. Evading the obligation of a king on coronation to swear to the constitution by not submitting himself to coronation at all (he had the Holy Crown conveyed to Vienna), Joseph pressed Hungary into his great machine. The counties were transformed into local branches of the state service, taking all their orders from above, and were required to conduct their business in German, which was also made the language of all education above the elementary level. The land was surveyed in preparation for taxing all estates in it equally. That the position of the peasants was improved pleased them but not their lords. When Joseph fell mortally sick, the country was on the brink of open revolt; and, although on his deathbed he retracted his administrative reforms, his successor, Leopold II (1790–92), was obliged to restore the ancient constitution and to swear to treat Hungary as a wholly independent kingdom, to be ruled only in accordance with its own laws and customs.

Francis I: the Reform Generation. When Leopold died with tragic suddenness in 1792, his young son, Francis, at first made the motions of conforming with his coronation oath, but he soon slipped back into the old ways. The Diet was convoked simply to supply money and, after 1811, was left unconvoked for 13 years. Social reaction accompanied the political absolutism, and the stranglehold on economic development was not relaxed.

For many years the Diet, composed either of magnates who identified their interests with those of the court, or of landowners who had prospered during the Napoleonic Wars, was as nonprogressive as Francis himself; but in wider circles the spirit of the age had given birth to a great cultural revival that was now bringing forth its first literary fruits, and the new national pride that it at once embodied and enhanced was demanding fulfillment of Leopold's promises and an end to the veiled but oppressive dictatorship of Vienna. A great reform movement was set on foot by Count István Széchenyi, who boldly proclaimed that the ancient privileges of the nobility were no bastion but a prison. He argued that the servile state of the peasants was humanly degrading and a source of weakness to the nation and also that the system of forced field labour, as well as the nobles' exemption from taxation, was economically harmful even to its supposed beneficiaries. After financial stringency had forced Francis to reconvoke the Diet in 1825, and thereafter to convoke it regularly, these doctrines were taken up, with variations, by a whole "reform generation," the most prominent figures of which were the legal expert Ferenc Deák; József, Baron Eötvös, leader of a little group of Centralists; and above all Lajos Kossuth, who largely changed the current of the reform movement by his insistence that social and economic reform could be fully realized only after the achievement of political independence. After Francis had been followed on the throne in 1835 by the luckless Ferdinand—in practice by the government of the two principal ministers, Klemens von Metternich and Anton von Kolowrat—Vienna was driven increasingly onto the defensive and forced to make repeated concessions, especially in respect of the replacement of Latin and German by Magyar as the language of the Diet, administration, and education—a demand pressed with especial insistence by many of the reformers.

Leaders of the reform movement

The nationalities. This raised a new and painful issue. The population of Hungary, even excluding Croatia, had never been purely Magyar, but the pre-Magyar inhabitants of the plains and the newcomers to them (outside the towns) had quickly become Magyarized; and, while this was not true of the peripheral areas, their populations were relatively sparse. By the end of the 15th century the Slovaks and Ruthenes of the north; the Germans of the free boroughs, the Szepes, and Transylvania; and the Romanians numbered hardly more than 20 to 25 percent of the total. The Magyar majority included almost the entire politically active "noble" class, the non-Magyar recruits to which assimilated most readily. The surviving non-Magyar peasants had neither the wish nor the ability to

question the Magyar character of the state, which for its part was uninterested in what languages were spoken by the politically disregarded, unfree men.

But between 1500 and 1800 the ethnic composition of the country changed. The most purely Magyar areas were heavily depopulated during the Turkish wars, and these losses were accompanied first by mass immigrations of Serbs, Croats, and Romanians from the Balkans and then by the introduction by the Austrian government of large numbers of German and other colonists. By 1720 the Magyars numbered only some 35 percent of the total population. By 1780 the figure had risen to nearly 40 percent, but the periphery, although it contained islands of Magyar population, was still preponderantly non-Magyar.

Further, ethnic conditions were no longer politically irrelevant. The Serb immigrants, who enjoyed cultural self-government under their patriarch, were bitterly anti-Magyar and anti-Hungarian. The resentment of the Romanians against their social grievances expressed itself in national terms. The other nationalities were slower to awaken, but among some of them, too, especially the Slovaks, there were nationalists who regarded the campaign to substitute Magyar for the neutral Latin as a threat to their national existence, while their resistance was interpreted by Magyar chauvinists as a threat to the security of Hungary, which, in their view, had to be politically unitary and Magyar in its superstructure. Some extremists pressed these demands with an intemperate impatience that at once justified and fed the very disaffection that it hoped to stifle.

From 1790 the Croats, too, were in conflict with Hungary. The first issue was a constitutional one: was Croatia in any way subordinate to Hungary, or was it Hungary's full equal? Then, when the linguistic campaign gained ground, the Croats consistently denounced as intolerable the suggestion that their representatives in the central services or in Parliament should be required to use Magyar instead of Latin. About 1830 a Croatian national movement emerged, with a political program that in its more extreme form demanded the reconstitution of Croatia, enlarged with other southern Slav areas, as an entirely separate unit within the Habsburg monarchy.

Revolution, reaction, and "compromise." The Hungarian reformers' opportunity came in the spring of 1848. Revolution of 1848 Inspired by the Revolution of 1848 in Paris, a popular upheaval caused the breakdown of central authority in Vienna. Two days later, on March 15 (a date celebrated in Hungary ever since), a bloodless revolution led by young intellectuals, including the poet Sándor Petőfi, abolished censorship in Pest and formulated the nation's demands. Seizing the moment, Kossuth prodded the Diet to rush through a body of laws (the April Laws), which—besides enacting important internal reforms, such as the generalizing of taxes, the abolition of villein status and the transfer of villein holdings to their cultivators, and the reorganizing of the lower "table" of Parliament on a representative basis—provided for the restoration of the territorial integrity of the lands of the Hungarian crown (subject, in the case of Transylvania, to the agreement of its Diet) and the appointment of a "responsible independent Hungarian Ministry." This was to include ministries of defense, finance, and external relations. A ministry of all the talents, headed by a progressive magnate, Count Lajos Batthyány, and including Kossuth (finance), Széchenyi (communications), Deák (justice), and Eötvös (education), took office. But the new government had enemies: the conservatives resented the land reform, and the Centralists in Vienna regarded the independent ministry, particularly its three "common" portfolios, as dangerous to the integrity of the monarchy. They found allies among the disaffected nationalities, notably the Serbs and Romanians, and in the Croats, whose ban, Josip Jelačić, refused to recognize the authority of Buda and Pest. Tension between Vienna and Pest mounted steadily, and in September, when the rest of the monarchy had been reduced, Jelačić, on Vienna's orders, invaded Hungary. Batthyány and other ministers resigned, leaving Kossuth in charge. An improvised national army drove Jelačić out of the country, but in December Ferdinand (whose coronation oath bound him to observe the April Laws) was made to abdicate in favour of his young nephew, Francis Joseph. The invasion was now renewed. A panmonarchic constitution abolished the April Laws, in reply to which a rump Diet, inspired by Kossuth, proclaimed the full independence of Hungary and the deposition of the Habsburg dynasty (April 14, 1849). The Hungarian forces, led by a young soldier of genius, Artúr Görgey, held their own until the Austrian court appealed for help to the Russian tsar, who sent an army across the Carpathians. Bitter fighting went on for some weeks more, but the odds were too heavy. On August 12, Kossuth fled the country, transferring his authority to Görgey, who the next day surrendered at Világos to the Russian commander.

Reprisals from Vienna Savage reprisals followed, and the country was again subjected to an absolutist and extortionate rule exercised from Vienna through a foreign bureaucracy. This "Bach regime" (after Alexander Bach, Austrian minister of the interior) was maintained, unrelaxed in principle although with some alterations in practice, until Austria's defeat in Italy in 1859 forced Francis Joseph to begin his retreat from absolutism. The followers of the exiled Kossuth were irreconcilable, but a large part of opinion inside Hungary rallied behind Deák. He held that the April Laws were legally valid and that Hungary's right to complete internal independence was inalienable but that under the Pragmatic Sanction, which he accepted, foreign affairs and defense were subjects "common" to the two halves of the monarchy and that a machinery could be devised for handling these affairs constitutionally. A Diet convoked in 1861 was dissolved after a few weeks, since the gap between the Hungarians' views and those of Francis Joseph and his Centralist ministry in Vienna proved to be still too wide to be bridged. Absolutism was reimposed, but the pressure of international and internal economic difficulties gradually drove Francis Joseph to see the necessity of further concessions. In July 1865 he dismissed his Centralist ministry; in December a new Diet was convoked and the negotiations reopened. Interrupted by the outbreak of the Seven Weeks' War, they were resumed after Austria's defeat by Prussia in 1866 had further convinced both parties of the necessity of agreement.

The Dual Monarchy, 1867–1918. A new Transylvanian Diet had already revoted the union with Hungary. In February 1867 Francis Joseph, having admitted the validity of the April Laws conditionally on the revision of those dealing with "common" affairs, appointed a responsible Hungarian ministry under Count Gyula Andrássy. A committee of the Diet then elaborated a law that, while laying down Hungary's full internal independence, provided for common ministries for foreign affairs and defense, each under a joint minister; a third common minister was in charge of the finance for these portfolios; these rendered account to delegations from the Austrian and Hungarian parliaments; the respective quotas to be paid for these services by each half of the monarchy were reconsidered every 10 years, as were commercial and customs agreements. At first, the two countries formed a customs union. On June 8 Francis Joseph was crowned king of Hungary, and on July 28 he gave his assent to the law.

Treatment of minorities Francis Joseph had stipulated that the settlement should include a revised Hungaro-Croatian agreement and provisions guaranteeing adequate rights for the non-Magyars of Hungary. The Croatian settlement (1868) left Croatia, including Slavonia, part of the Hungarian crown, under a ban appointed on the proposal of the Hungarian prime minister. Croatia was to enjoy full internal autonomy, but certain questions were designated as common to Croatia and Hungary. When these were under discussion, Croatian deputies attended the central Parliament, in which they could speak in Croat. The sole language of internal official usage in Croatia was Croat.

The Nationalities Law (1868) laid down that all citizens of Hungary, whatever their nationality, constituted politically "a single nation, the indivisible, unitary Hungarian nation," and there could be no differentiation between them except in respect of the official usage of the current languages, and then only insofar as necessitated by practical considerations. The language of the central admin-

istrative and judicial services, and of the university, was Magyar, but there were to be adequate provisions for the use of non-Magyar languages on lower levels. The consolidation was completed by the incorporation of the military frontier (in stages lasting several years) and of Transylvania, the latter process involving the abolition of the old "Three Nations," except that the Saxon "university" was allowed to survive as a purely cultural institution.

Hungary under dualism. The Compromise (Ausgleich) of 1867 restored territorial integrity to Hungary and gave it more real internal independence than it had enjoyed since 1526; the monarch's powers in internal affairs were strictly limited. In the conduct of foreign affairs or defense, however, Hungary still formed only part of the monarchy, and its interests in these fields had to be coordinated with those of its other components. But Hungary had a large voice in the monarchy's policy in these fields and enjoyed the great advantage—which weighed heavily with soberer men, including Deák, when negotiating the Compromise—that the resources of the great power of which it formed a part stood behind the country. To some, however, the price still seemed too high, and the parliamentary life of Hungary from 1867 to 1918 was dominated by the conflict between the supporters and the opponents of the Compromise, the latter ranging from complete separatists to those who accepted the Compromise in theory but wanted details of it altered.

The supporters of the Compromise, then known as the Deák Party, were in office first but soon got into such financial and personal difficulties that complete chaos threatened. It was averted when in 1875 Kálmán Tisza, the leader of the moderate nationalist Left Centre, carried through a fusion between his party and the remnants of the Deákists on a program that amounted to putting his party's main demands into cold storage until the political and financial situation was stabilized. This new Liberal Party then held office for nearly 30 years. During these years the Compromise stood intact, but there was mounting friction with Vienna over the army, which the Hungarians regarded, with some reason, as imbued with a spirit hostile to themselves; over the economic provisions of the Compromise; and over the question of Hungarian participation in control of the National Bank. An army question in 1889 marked something of a turning point, after which relations between the supporters of the Compromise, behind whom stood the crown, and its nationalist opponents were permanently strained. The tension reached a climax in 1903, when the obstruction of the "national opposition" made parliamentary government practically impossible. The premier, Count István Tisza (Kálmán Tisza's son), dissolved Parliament. Elections in January 1905 gave a coalition of "national" parties a parliamentary majority, but Francis Joseph refused to entrust the government to them on their program, which included "national" concessions over the army, and a period of nonparliamentary government followed until in April 1906 the coalition leaders, under threat of an extension of the suffrage if they proved recalcitrant, gave the king a secret undertaking that, if appointed, they would not press the essentials of their program. On this basis he appointed a coalition government, but under a Liberal, Sándor Wekerle. With their hands thus tied, the coalition made a wretched showing. Tisza reorganized the Liberal Party as the "Party of Work," and in the elections of 1910 this secured a large majority. After Count Károly Khuen-Héderváry (1910–12) and László Lukács (1912–13), Tisza himself again became prime minister, and Francis Joseph ceased to press his demand for effective franchise reform, to which Tisza was inexorably opposed.

The rise of the Liberal Party

Social and economic developments. Hungary underwent much change after 1867. The achievements of the Deákist and Liberal governments included, besides the assimilation of the former outlying areas of Transylvania and the military frontier, a reform of the relations between the central government and the counties, and a general reorganization of the administration. The judicial system was modernized. Relations between the state and the churches were, after a long struggle, restated in 1894–95 on terms satisfactory to the liberal philosophy of the day, which brought with them full emancipation for Hungary's large Jewish population. In 1868 Eötvös carried through an admirable elementary education act, and much headway was made in raising the educational and cultural level of the country. After long difficulties the national finances were put in order and the public debt reduced. There was considerable economic progress in many fields. Agriculture remained the mainstay of the economy. The medium and small landowners had been hard hit by the land reform of 1848, but the survivors were helped by the high agricultural prices and the secure Austrian market. Afterward, the general European agricultural depression plunged even the big landowners into difficulties, but these diminished near the turn of the century, when prices rose again, while the quality and quantity of production improved. Many branches of industry failed to survive the customs union with Austria, but the agricultural industries prospered, and later, as domestic capital accumulated, a process of industrialization, helped by state legislation, set in which became extremely rapid after 1890. As late as 1910 agriculture was still the most important branch of the economy, and 70 percent of the population still derived its livelihood from the soil, but more than 16 percent did so from industry and mining.

Economic progress

Urbanization proceeded apace. The growth of Budapest was meteoric—its population before World War I topped 800,000—and two other cities had populations of more than 75,000 and a dozen around 50,000. Communications were largely modernized.

For all this, Hungary was still a relatively poor country. The extremely rapid growth of the population (from 13 million in 1850 to more than 20 million in 1910) had far outstripped that of the means of production. The growth of industry was still too slow to absorb the surplus rural population, and in spite of high emigration, which in the last years averaged 100,000 annually, acute rural congestion had developed. While 35 percent of the land was held in 4,000 large estates, there were about two million small, or dwarf, holdings, and a further 1.7 million persons (wage earners) were totally landless. A large proportion of this rural proletariat was forced to live in conditions of extreme misery, near starvation. The living standards and conditions of the industrial workers, especially the unskilled, were also very low.

The political structure was not modern. The unreformed franchise excluded the masses from political influence, and even the vocational organization that they were able to achieve was primitive. The industrial and financial development had been largely the work of Jews (who also held a large part in the professions) or of Magyarized Germans. Its own quasi-alien character and its small numbers prevented the Hungarian bourgeoisie from developing into a positive factor in the political life, which continued to be dominated by a landowning class whose social and political ideas failed to move with the times.

The "nationalities problem" remained intractable. After 1868 Hungarian political philosophy insisted more strongly than ever that the Hungarian state must be Magyar in spirit, in its institutions, and, as far as possible, linguistically. Suggestions to the contrary, or appeals to the Nationalities Law, met with derision or abuse. In spite of the law, the use of minority languages was banished almost entirely from administration and even justice. While the autonomy of the church schools was hardly attacked until the 20th century, most denominations saw to it that all secondary education in their schools, with trivial exceptions, should be in Magyar, which was also represented far above its due in the primary schools, as was practically all instruction in the state schools founded from 1870 onward.

By the end of the century the state apparatus was entirely Magyar in language, as were business and social life above the lowest levels. The proportion of the population with Magyar as its mother tongue rose from 46.6 percent in 1880 to 51.4 percent in 1900. The Magyarization of the towns had proceeded at an astounding rate. Nearly all middle-class Jews and Germans and many middle-class Slovaks and Ruthenes had Magyarized.

Magyarization of the middle classes

Most of the Magyarization, however, had been in the

centre of Hungary and among the middle classes. It had hardly touched the rural populations of the periphery, and the linguistic frontiers had hardly shifted from the line on which they had been stablilized in the 18th century. In these areas, moreover, a hard core of national feeling had survived. This had weakened during the first decades after the Compromise but was reviving again at the turn of the century, especially among the Romanians, and was being encouraged from across the frontiers of Romania and Serbia and (in the case of the Slovaks) from Bohemia. Hungaro-Croatian relations, too, deteriorated, after a period of quiescence, when the Serbian government began propagating a theory of "Yugoslav" unity designed to detach the Croats from the monarchy.

Many of these developments threatened the very basis of the Compromise, and to this another uncertainty was added. Francis Joseph could be trusted to support and accept the policies of any Hungarian government that on its side maintained the Compromise loyally; but he was an old man, and his heir presumptive, the archduke Francis Ferdinand, was notoriously hostile to the Hungarian regime. In touch with many of its opponents, the archduke was credited with designs of overthrowing the Compromise to the benefit not of its traditional opponents, the Hungarian Independents, but of its enemies in the opposite camps, especially the nationalities.

World War I. The assassination of the archduke on June 28, 1914, removed this danger and plunged Austria-Hungary into World War I. For the first two years of the war, Tisza upheld the internal system and held the country to its international course; and when Francis Joseph died he persuaded the new king, Charles IV, to accept coronation (Dec. 30, 1916), thus binding himself to uphold the integrity and constitution of Hungary. Charles, however, insisted on electoral reform, and Tisza resigned (May 1917). While short-lived minority governments struggled with increasing difficulties, the threefold agitation grew: of the Hungarian nationalists, against a war into which, they maintained, Hungary had been drawn in the interest of Germany and Austria; of the political left, growing daily more radical under the stimuli of privation and the Russian Revolution; and of the nationalities, encouraged by the favour that their kinsfolk were finding with the Triple Entente. The country began to listen to Count Mihály Károlyi, leader of a faction of the Party of Independence, who proclaimed that a program of independence from Austria, repudiation of the alliance with Germany, and peace with the Entente, combined with social and internal political reform and concessions to the nationalities, would safeguard Hungary against all dangers at once.

REVOLUTION, COUNTERREVOLUTION, AND THE REGENCY, 1918–45

On Oct. 31, 1918, when the defeat of the monarchy was imminent, Charles appointed Károlyi prime minister at the head of an improvised administration based on a left-wing National Council, and, after the monarchy had signed an armistice on November 3 and Charles had "renounced participation" in public affairs on the 13th, the National Council on the 16th dissolved Parliament and proclaimed Hungary an independent republic, with Károlyi as provisional president. The separation from Austria was popular, but all Károlyi's supposed friends disappointed him, and all his premises proved mistaken. Serb, Czech, and Romanian troops installed themselves in two-thirds of the helpless country, and, in the confusion, ordered social reform was impossible. The government steadily moved leftward and on March 21, 1919, Károlyi's government was replaced by a soviet republic, controlled by Béla Kun, who had promised Hungary Russian support against the Romanians. The help never arrived, and Kun's doctrinaire bolshevism, resting on a "red terror," antagonized almost the entire population. On August 4 Kun and his associates fled, and two days later Romanian troops entered Budapest. Shadow counterrevolutionary governments had already formed themselves in Szeged (then occupied by French troops) and Vienna and pressed the Allies to entrust them with the new government. The Allies insisted on the formation of a provisional regime including democratic elements that must hold elections on a wide, secret suffrage. The Romanians were, with difficulty, induced to retire across the Tisza; and a government, under the presidency of Károly Huszár, was formed in November 1919 and the elections (for a single house) held in January 1920. The new Parliament began by declaring all measures enacted by the Károlyi and Kun regimes, and also the legislation embodying the Compromise of 1867, null and void. The institution of the monarchy was thus restored, but, pending a settlement, on which feelings were deeply divided, of the relationship between the nation and the dynasty, Admiral Miklós Horthy, who had organized the counterrevolutionary armed forces, was elected regent as provisional head of state (March 1, 1920). The Huszár government then resigned, and on March 14 a coalition government, composed of the two main parties in the Parliament (the Christian National Union and the Smallholders), took office under Sándor Simonyi-Semadam.

The soviet republic of Béla Kun

The Regency, 1920–45. *The Treaty of Trianon.* The Allies had long had their peace terms for Hungary ready but had been unwilling to present them to an earlier regime. It was, thus, the Simonyi-Semadam government that was forced to sign the Treaty of Trianon (June 4, 1920). This was hard indeed, for the Allies not only assumed without question that Hungary's non-Magyar populations should be deemed to be wishing to leave Hungary but also allowed the successor states, especially Czechoslovakia, to annex large areas of Magyar population. The final result was to leave Hungary with only 35,893 square miles of the 125,641 square miles that had constituted the lands of the Hungarian crown. Romania, Czechoslovakia, and Yugoslavia all took large fragments, while others went to Austria and even to Poland and Italy. Of the population of 20,866,447 (1910 census), Hungary was left with 7,615,117. Romania received 5,257,467; Czechoslovakia, 3,517,568; Yugoslavia, 4,131,249; and Austria, 291,618. Of the 10,050,575 persons for whom Magyar was the mother tongue, no fewer than 3,219,579 were allotted to the successor states: 1,704,851 to Romania, 1,063,020 to Czechoslovakia, 547,735 to Yugoslavia, and 26,183 to Austria. While the homes of some of these—*e.g.*, the Szeklers—had been in the remotest corners of historic Hungary, many were living immediately across the frontiers.

In addition, the treaty required Hungary to pay in reparations an unspecified sum, which was to be "the first charge upon all its assets and revenues," and limited its armed forces to 35,000, to be used exclusively for the maintenance of internal order and frontier defense.

Postwar confusion and reconstruction. Conditions in Hungary in 1920 were exceedingly difficult in every respect. The prolonged war, the Bolshevik regime (before which mobile capital had fled headlong), and the rapacious Romanian occupation had exhausted its resources, and the economy had been further disrupted by the new frontiers, which cut the factories off from both their accustomed supply sources and their markets. Industrial unemployment had reached unprecedented heights, and the surviving national resources were being strained to support nearly 400,000 refugees from the successor states.

Both the industrial and the agrarian proletariats were embittered by the failure of their revolutionary hopes. Even more dangerous were the armies of the "new poor," not only the homeless refugees but also a large part of the middle classes in general, reduced to penury by the galloping inflation. They formed another radical army, but its radicalism was one of the right: they ascribed their misery precisely to the revolutions, on which they put the blame for all Hungary's misfortunes. Feeling ran particularly high against the Jews, who had played a disproportionately large part in both revolutions, especially Kun's, but the resentment extended also to the Social Democrats and even to Liberal democracy. "White terrorists" wreaked indiscriminate vengeance on persons whom they associated with the revolutions. Huszár's government itself had turned so sharply on the Social Democrats and the trade unions that the former withdrew their representatives from the government and boycotted the elections, in protest against the widespread arrests and internments.

The Right Radicals

The government of Pál Teleki, who succeeded Simonyi-

Semadam in July 1920, blunted the edge of the agrarian unrest by a modest reform—promised, indeed, only as a first installment—that took 1.7 million acres (7.5 percent of the total area of the country) from the biggest estates for distribution in smallholdings, but it had hardly touched any other social problem when in March 1921 the legitimist question was raised in acute form by King Charles' sudden return to Hungary. He was forced to withdraw, the command coming from the Allies but too willingly obeyed by the Right Radicals, toward whom Horthy was then leaning. The government, several of whose members were legitimists, resigned, and the succession was assumed by the conservative Count István Bethlen, who had been waiting behind the scenes for his cue. Bethlen devised a formula that, while not legally excluding the king's return (under Entente pressure, Parliament had voted a law dethroning the Habsburgs, but even Hungary's own anti-legitimists never took it as morally binding), excluded it in practice. In return for this, the Smallholders agreed with the antilegitimists of the Christian Nationals to form a new Party of Unity under Bethlen's leadership. He persuaded Parliament to accept as still legally in force the franchise enacted in 1918, which reduced the number of voters and reintroduced open voting in rural districts. Elections in May 1922 gave this party a large majority.

The end of legitimism

Meanwhile, a second attempt by the king (in October 1921) to recover his throne again ended in failure, and soon after, the legitimist question lost its acuteness with Charles' death. In December 1921 Bethlen concluded a secret pact with the Social Democrats, under which the latter promised to abstain from agitation on the land and to support the government's foreign policy in return for the cessation of persecution, the release of political prisoners, and the restoration of the sequestrated trade union funds. The peasant leaders were persuaded to accept the indefinite postponement of the second installment of the land reform. The white terror was liquidated quietly but effectively, chiefly by finding government employment for the Right Radical leaders.

Bethlen's domestic program was made possible by his cautious international policy. Almost all Hungarians were passionately convinced of the injustice of the Treaty of Trianon, revision of which was the all-dominant motive of Hungary's foreign policy throughout the interwar period and the key to the hostile relations between Hungary and those states that had chiefly profited by it. Bethlen was as revisionist at heart as any of his countrymen, but he was convinced that Hungary could not act effectively in this field until it had acquired friends abroad and had achieved political and economic consolidation at home. This depended on financial reconstruction; and, to achieve this, he applied for Hungary's admission to the League of Nations, which was granted (not without difficulty) in September 1922. In March 1924, in return for an undertaking to carry out loyally the obligations of the treaty, he obtained a League loan, which had almost magical effects. The inflation stopped immediately; the League loan was followed by a flood of private lending; and the expatriated domestic capital returned. With this help, Hungary enjoyed some years of prosperity, during which agriculture revived and industrialization made progress.

Abroad, Bethlen's only other important move was the conclusion in 1927 of a treaty of friendship with Italy. At home his regime, which was conservative but not tyrannical, rested on what came to be called Hungary's conservative-liberal forces, to the exclusion of extremism from left or right.

Financial crisis: the rise of right radicalism. Bethlen's command of Parliament was complete and was not even shaken by the disastrous fall in world wheat prices in 1929. In June 1931 he had just held elections that returned his party with the conventional large majority when the supervention of a world financial crisis on the economic one shattered the foundations of his structure. The foreign creditors called in their money, and Hungary, its trade balance annihilated by the collapse of the wheat market, could not meet their demands and had to apply for help from the League of Nations, which imposed a regime of rigid orthodox deflation. Industrial unemployment soared again, the agricultural population was rendered almost literally penniless, and the government services were forced to carry through large-scale dismissals and salary reductions in the interests of a balanced budget.

In August Bethlen resigned. His successor, Count Gyula Károlyi, was unable to cope with the situation. Political agitation mounted, and on Oct. 1, 1932, Horthy appointed as prime minister the leader of the Right Radicals, Captain Gyula Gömbös.

Gömbös appointed prime minister

At home Gömbös found the leading strings of financial forces, international and domestic, as invincible as had his predecessors. Previously a violent anti-Semite, he had to make a public recantation of his views on this point and was unable to carry through any other points of his neofascist program, particularly as Horthy at first refused to allow him to hold elections. Neither was he able to realize his foreign political ideal of an "Axis" composed of Hungary, Italy, and Germany, since his two proposed partners were then at loggerheads over Austria, and Gömbös, one of whose first acts had been to dash to Rome and breathe new life into Hungary's friendship with Italy, found himself drawn into the "Rome Triangle" (Italy, Austria, and Hungary) that was directed precisely against Germany. Finally, Hitler upset another of his calculations by telling him that, while Germany would help Hungary against Czechoslovakia, it would not do so against Romania or Yugoslavia.

Nevertheless, when Gömbös died prematurely in October 1936, he could feel that his years of office had not been entirely wasted. Shortly before, Horthy had at last allowed him to hold elections, and these had brought into Parliament a strong Right Radical contingent from which it could never thereafter free itself. Abroad, when Benito Mussolini surrendered to Hitler, Hungary found itself after all in a sort of Axis camp, membership of which might help it at least to accomplish partial revision of the Treaty of Trianon. On the other hand, if Germany chose to apply economic or political pressure, Hungary would be defenseless, but for such shadow help as Italy could offer.

This threat was already looming large, and thenceforward it became inextricably involved with Hungary's own internal politics, by reason of the ideological character of the Nazi regime and in particular its anti-Semitism. Anti-Semitism at that stage was running high in Hungary itself, and those infected by it—the Right Radicals of various brands, but even other members of the middle classes—welcomed Germany's support for their own ideas while making light of its dangers. They even argued, not without reason, that the danger lay in affronting Germany, which could easily crush unarmed little Hungary but would not wish to attack a friend and ideological partner. They further believed (nearly all the army officers held this view) that, should Hitler's policies lead to war, Germany would emerge the victor; Hungary's salvation thus lay in joining forces with Germany. On the other side there came into being a curious shadow front, composed of all elements antagonistic to Nazism—not only Hungary's Jews but also the legitimists, the traditionalist conservative-liberals, and the Social Democrats. Many of these were not convinced that Germany was invincible, and they held that, if war came, only disaster could follow for Hungary if it became too closely involved with Germany. Even they, however, were unwilling to draw the ultimate conclusion that Hungary should abandon all its revisionist claims and join hands with the Little Entente, which, for its part, indicated that it would accept nothing short of total renunciation. It was of the highest importance that by this time Horthy had shed his earlier Right Radical leanings and sympathized with this shadow front.

The question of support for Germany

To succeed Gömbös, Horthy appointed Kálmán Darányi, who was more of a conservative than a Right Radical; the appointment was ill received in Germany, which grew even more hostile the next year, when Darányi's foreign minister, Kálmán Kánya, obtained the tacit consent of the Little Entente for Hungary to rearm, although Hungary was still sadly short of armaments, for which, again, Germany was its only source of supply. On a visit to Berlin, Darányi and Kánya smoothed over the difficulties; but, when Darányi tried to placate the extremists

at home, Horthy replaced him (in May 1938) with Béla Imrédy, who introduced a largely token "Jewish Law" but nevertheless pinned his hopes on the West. When the Munich crisis broke in September, Imrédy and Kánya, while presenting Hungary's claims on Czechoslovakia, limited the claims to what they hoped would be acceptable to the Western powers, whose endorsement they made every effort to obtain. The West ignored them, however, and the Hungarian leaders had, after all, to turn to Germany and Italy, which, under the "First Vienna Award" of November 2, gave Hungary the Magyar-inhabited fringe of southern Slovakia. Imrédy, disillusioned with the West, dismissed Kánya for the pro-Axis Count István Csáky and sought to recover Hitler's favour by introducing a more far-reaching (anti-)Jewish Law. Imrédy's enemies secured his resignation in February 1939 by unearthing documents purporting to show a Jewish strain in his own ancestry. Pál Teleki, who succeeded him, was a most convinced Westerner; but Hungary's recovery of Carpatho-Ruthenia (March 1939) was, again, sanctioned by Hitler.

War and renewed defeat. When Germany attacked Poland (Sept. 1, 1939), Hungary refused to allow German troops to cross Hungarian territory. In the first months of World War II none of the belligerents wanted the war extended to southeastern Europe, and so Teleki and Horthy were able to keep Hungary nonbelligerent and, after the Soviet Union had occupied Bessarabia in June 1940, to force a reluctant Germany (but a willing Italy) to cede to it northern Transylvania under the "Second Vienna Award" of August 30. But they then allowed German troops to cross Hungarian territory into southern Romania, and in November they signed the Tripartite Pact.

German troops in Hungary

The next step was more fatal still. In his search for cautious reinsurance, Teleki concluded with the like-minded government of Yugoslavia a treaty (December 12), unluckily named one of "Eternal Friendship." On March 26, 1941, that Yugoslav government was overthrown by a pronouncedly pro-Western one. Hitler prepared to invade Yugoslavia and called on Hungary to help him. Hungary, caught in an unanticipated situation, refused to join in the attack but again allowed German troops to cross its territory. Great Britain threatened to declare war; and Teleki, blaming himself for having allowed to come into being a situation that it had been his life's aim to avoid, took his own life on April 2. His successor, László Bárdossy, waited until Croatia had declared its independence (April 10); then, arguing that Yugoslavia had already disintegrated, he occupied the ex-Hungarian areas of Yugoslavia.

Bárdossy, although no fascist, believed that the Axis would win the war and that Hungary's salvation lay in placating it; otherwise, he believed, Romania (now pro-Axis) would persuade Hitler to reverse the Second Vienna Award. Accordingly, when Germany attacked the Soviet Union (June 22), Bárdossy sent a token force to assist in what everyone expected to be a brief operation. The strength of the Soviet resistance upset the calculation, and in January 1942 the Germans forced Hungary to mobilize practically all its available manpower and send it to the Soviet Union. Meanwhile, Britain had identified the Western cause with that of the Soviet Union. In December 1941 Britain declared war on Hungary, which in its turn declared war on the United States. Further, Britain recognized the Czechoslovak government in exile and withdrew recognition of the First Vienna Award, and the Soviet Union recognized Czechoslovakia's 1937 frontiers.

Hungary in the war

Many Hungarians by then agreed with Bárdossy that Hungary's only course was to fight on until the Axis won the war—the more so because all Hungarians except those of the extreme left regarded bolshevism as the embodiment of evil. Horthy, however, while sharing the latter view, still believed in a Western victory and thought it possible for Hungary, while continuing the struggle in the east, to regain the favour of the West. In March 1942 he replaced Bárdossy with Miklós Kállay, who shared these hopes. For two years Kállay conducted a remarkable balancing act—protecting Hungary's Jews and allowing the left (except for the communists) almost untrammeled freedom while putting out innumerable feelers to the West, to which he actually promised to surrender unconditionally when its troops reached Hungary's frontiers. Meanwhile, in January 1943 the Hungarian expeditionary force suffered a crushing defeat at Voronezh that cost it much of its manpower and nearly all its equipment. (C.A.M./Ed.)

But the Western forces did not approach the Danube valley; and, as the Soviet army neared the Carpathians, Hitler, from whom few of Kállay's activities were hidden, decided that he could not leave his vital communications at the mercy of an untrustworthy regime. In March 1944 he summoned Horthy and offered him the choice between full cooperation under German supervision or undisguised German occupation with the treatment accorded to an enemy country. Horthy chose the former course and appointed a collaborationist government under General Döme Sztójay. For a while the Germans did much as they wished—they suppressed parties and organizations of potential opponents and arrested their leaders. With the cooperation of Hungarian authorities, Jews were compelled to wear a yellow star, robbed of their property, and incarcerated in ghettos as in other Nazi-occupied areas. Except for the Jews in the capital and those in the forced-labour camps of the Hungarian army—whose turn would come later—they were deported to the gas chambers of German concentration camps. In spite of the efforts of representatives of some neutral countries—such as Raoul Wallenberg of Sweden, the papal nuncio, and diplomats of Switzerland, Portugal, and even Spain, who saved tens of thousands of lives—more than 550,000 of Hungary's nearly 750,000 Jews, as defined by racist legislation, perished during the war.

In the summer the pressure relaxed; and in August, after Romania's surrender to the Allies, Horthy appointed a new government under the loyal general Géza Lakatos and reopened the peace feelers. A mission was sent to Moscow, where it duly concluded a "preliminary armistice"; but, when on October 15 Horthy announced this on the wireless, he was abducted by the Germans, who forced him to recant and to abdicate. The Germans put Ferenc Szálasi, the leader of the right-wing extremist Arrow Cross movement, in charge. By then, however, Soviet troops were far inside the country. The Germans and their Hungarian allies were driven slowly back, while numerous refugees fled with them. The last armed forces crossed the Austrian frontier on April 4, 1945.

Soviet troops in Hungary

The defeat was sealed in a new peace treaty, signed in Paris on Feb. 10, 1947, which restored the Trianon frontiers, with a rectification in favour of Czechoslovakia and the Soviet Union. It imposed on Hungary a reparations bill of $300 million and limited its armed forces. The implementation of the treaty's provisions was to be supervised by a Soviet occupation force, a large contingent of which remained in the country.

THE PEOPLE'S REPUBLIC

As in 1920, a new regime recognized the defeat of its predecessor. As early as Dec. 31, 1944, a makeshift Provisional National Assembly had accepted a government list and program presented to it by communist agents following in the train of the Soviet armies. The communists began cautiously. They announced that the new Hungary was to rest on "all its democratic elements." The government contained only two communists; its other members were representatives of four noncommunist left-wing parties—the Smallholders, the Social Democrats, the National Peasants, and the Progressive Bourgeoisie—and four men associated with the Horthy regime, including two generals who had been in Moscow in connection with the armistice talks. The program provided for the expropriation of the large estates and the nationalization of the banks and heavy industry; but it promised guarantees of democratic rights and liberties, respect for private property as such, and encouragement of private initiative in trade and small industry.

The communist regime. The full political takeover, however, proceeded systematically, although not according to any timetable, for the communists, misjudging feeling in the country, allowed the first elections (November 1945) to be relatively free. Only the parties of the coalition were allowed to contest them; but the adherents of the pro-

scribed parties voted for the Smallholders, who received an absolute majority. The head of the Soviet mission, however, insisted that the coalition must be maintained; a Smallholder was allowed to be prime minister, but the Ministry of the Interior, with the control of the police, was given to the communists. Pressure and intimidation were then applied to the Smallholders to expel their more courageous elements as "fascists"; and in the next elections (August 1947) the Smallholders polled only 15 percent of the votes cast. The communists had meanwhile forced the Social Democrats to form a "workers' bloc" with them. Although the pressure this time was considerable, the bloc still polled only 45 percent of the votes (other parties were allowed to stand this time); but the communists then forced the Social Democrats to fuse with them in a single "Workers' Party," from which the recalcitrants were expelled.

Communist takeover

In the next elections (May 1949) voting was open, and the voters were presented with a single list, on which Smallholders and National Peasants still figured, but the persons using these descriptions were mere tools of the communists. In August a new constitution was enacted—a copy of that of the Soviet Union. Hungary, a republic since Feb. 1, 1946, now became a "people's republic," and, although its president (Zoltán Tildy) and for a while its prime ministers (Ferenc Nagy, then Lajos Dinnyés) were Smallholders, all real power rested with the Workers' Party, which was controlled by its first secretary, Mátyás Rákosi, by then the real power in the country. Finally, the party's "Muscovite wing" turned on its "national wing." The leader of this latter group, László Rajk, was executed on questionable charges in October 1949, and his chief adherents were similarly executed or imprisoned. Meanwhile, hundreds of persons had been executed or imprisoned as war criminals, many of them for no other offense than loyalty to the Horthy regime. Many thousands more were interned. The State Security Department (AVO) was omnipotent. The judiciary, civil service, and army were purged, and party orthodoxy became the criterion for positions in them. The trade unions were made into mere executants of party orders. After the dissolution of the parties, the chief ideological opposition to the communist regime came from the churches; but their estates were expropriated, making it impossible for them to maintain their schools, and in July 1948 the entire educational system was nationalized. The Calvinist and Lutheran churches accepted arrangements under which the state made good their remaining expenditure. The head of the Roman Catholic church, József Cardinal Mindszenty, who refused to follow their example, was arrested on transparent charges in December 1948 and condemned to life imprisonment. The monastic orders were dissolved. After this the Roman Catholic church accepted financial terms similar to those of the other churches, and eventually the bishops, with visible repugnance, took the oath of loyalty to the state.

The communists' economic program, like their political program, could not be realized immediately, because in 1945 the country was in a state of economic chaos worse even than that of 1918. This time the country had been a theatre of war. Many cities, notably Buda, were in ruins, and communications were wrecked; the retreating Germans had destroyed the bridges between Buda and Pest and had taken with them all they could of the country's portable wealth. The Soviet armies lived off the land, and the Soviet Union took its share of reparations in kind, placing its own values on the objects seized. It also took over former German assets in Hungary, including Jewish property confiscated during the Nazi occupation.

Economic programs

A three-year plan introduced in August 1947 was devoted chiefly to the repair of immediate damage. This was declared completed, ahead of schedule, on Dec. 31, 1949. By then the communists were in full political control, and measures nationalizing banking, most industry, and most internal and all foreign trade had been enacted. The land, outside the big estates, was not touched at first, but in 1948 Rákosi announced a policy of collectivization of agriculture. Three forms were envisaged: state farms and two types of cooperative. Peasants were forced by various pressures into the cooperatives, the character of which approached ever more closely that of the state farms.

The three-year plan was succeeded by a five-year plan, the aim of which was to turn Hungary into a predominantly industrial country, with heavy industry taking first place. Huge sums were devoted to the construction of foundries and factories, many of them planned with little regard for Hungary's real resources and less still for its needs. In fact, the plan was concerned with the needs of the Soviet Union, for which Hungary was to serve as a workshop. Hungary's newly discovered deposits of uranium went straight out of the country. Industrial production rose steeply, but the standard of living did not; the production of consumer goods was throttled and that of agriculture stagnated.

The revolution of 1956. Rákosi—who in 1952 came to preside over the government as well as the party—was, under Moscow, all-powerful until the death of Stalin in 1953, when a period of fluctuations set in. In July 1953 Rákosi was deposed from the prime ministership in favour of Imre Nagy—a "Muscovite" but a Hungarian in his attitudes and not unpopular in the country. Nagy promised a new course—end of the forced development of heavy industry, more consumer goods, no more forcing of peasants into the collectives, release of political prisoners, and closing of internment camps. He introduced some of these reforms, but Moscow hesitated to support him. In the spring of 1955 Nagy was dismissed from office and expelled from the party. Rákosi, reinstated, put the country back on its previous course but was dismissed again in July 1956, this time from all his offices and in disgrace. The new Soviet leader, Nikita S. Khrushchev, had sacrificed Rákosi as a gesture to President Tito of Yugoslavia, whom he wished to placate and whom Rákosi had offended personally. The new man in charge was Ernő Gerő, Rákosi's deputy and almost as detested as Rákosi himself. Gerő promptly announced that there would be no concessions on matters of principle to Nagy and his group.

The relaxation of pressure under Nagy, however (though transitory), Khrushchev's "secret speech" denouncing Stalin's cult of personality delivered at the 20th Congress of the Communist Party of the Soviet Union (February 1956), and the Polish challenge to the Soviet Union in the spring and summer of 1956 emboldened Hungarians. On October 23, students in Budapest staged a great procession, which was to end with the presentation of a petition asking for redress of the nation's grievances. People flocked into the streets to join them. Gerő answered with an unwise and truculent speech, and police fired into the crowds. The shots turned a peaceful demonstration into a revolutionary one. The army joined the revolutionaries, and army depots and munitions factories handed out arms. Outside Budapest local councils sprang up in every centre; the peasants reoccupied their confiscated fields. The communist bureaucracy melted away. Prison doors were opened. The members of the AVO fled if they could. A cheering crowd escorted Cardinal Mindszenty back to the palace.

Revolution and change

In kaleidoscopic political changes, Nagy resumed power but was driven from one concession to the next, until he found himself at the head of a genuine coalition government composed of Smallholders, Social Democrats, and National Peasants, which, with a "Catholic Association," had reconstituted themselves.

The Soviet troops had withdrawn, and Nagy was negotiating for the complete evacuation of Hungary. On November 1 he announced Hungary's withdrawal from the Warsaw Pact (to which it had adhered since 1955) and asked the United Nations to recognize Hungary as a neutral state, under the joint protection of the Great Powers. High officials from Moscow flew to Budapest and were in two minds whether to let matters take their course. But Nagy's denunciation of the Warsaw Pact seemed too dangerous to them, and their tanks, which had halted just across the frontier, began to return, reinforced by other units. By November 3, the tanks were in position around the main centres of Hungary; at 4 AM on the 4th they entered Budapest. Nagy took refuge in the Yugoslav Embassy and Cardinal Mindszenty in the U.S. Legation. General Pál Maléter, head of the Hungarian national forces, who

had been invited by the Soviet commanders to negotiate, was imprisoned.

A communist leader, Ferenc Münnich, speaking from a radio station behind the Soviet lines, and János Kádár, a "National Communist" who had been imprisoned under Rákosi and had actually joined the revolutionaries on October 24, formed a new "revolutionary peasant-worker government," consisting entirely of communists, with Kádár as prime minister. Kádár promised that when the "counterrevolution" had been suppressed and order restored he would negotiate on the withdrawal of the Soviet garrison (although the denunciation of the Warsaw Pact was retracted); he dissociated himself from the "Rákosi–Gerő clique" and promised internal reforms.

The country was not convinced, and fighting broke out. But the odds were too heavy, and the major hostilities were over within a fortnight. The workers, however, proclaimed a general strike, and it was many weeks before they were brought to heel.

The end of the revolution

Meanwhile, Nagy, who had left his place of refuge under safe conduct, had been abducted to Romania. After a secret trial he, Maléter, and a few close associates were executed in 1958. Many lesser figures were seized and transported to the Soviet Union, some never to return; 200,000 refugees escaped to the West. Thus, a substantial proportion of Hungary's educated classes was lost to the country. Material damage was also heavy.

The Kádár regime. In the first uncertain weeks of his regime Kádár made many promises. Workers' councils were to be given a large amount of control in the factories and mines. Compulsory deliveries of farm produce were to be abolished, and no compulsion, direct or indirect, was to be put on the peasants to enter the collectives. The five-year plan was to be revised to permit more production of consumer goods. The exchange rate of the ruble and forint was to be adjusted and the uranium contract revised. For a time there was even talk of a coalition government.

The larger hopes were dashed after representatives of the Soviet Union, East Germany, Czechoslovakia, Romania, and Bulgaria conferred with those of Hungary in Budapest in January 1957. A new program was soon issued stating that Hungary was a dictatorship of the proletariat, which in foreign policy relied on the Soviet Union and the Soviet bloc. Further, it was asserted that the Soviet garrison was in Hungary to protect the nation from imperialist aggression. Internal reforms were again promised, however, and foreign trade agreements were to be based on complete equality and mutual advantage.

Subsequently, Kádár was at great pains to give the Soviet Union no cause for uneasiness over Hungary's loyalty. Whenever any international issue arose, he invariably supported Moscow's policy with meticulous orthodoxy, even sending a contingent into Czechoslovakia in 1968. At home he ignored some of his promises and honoured others only superficially. The peasants were put under such pressure to enter cooperatives that within a few years practically no private farms survived. The workers' councils were dissolved, but trade unions were later granted rights to query decisions by management. Parliament remained a rubber stamp, and a Patriotic People's Front (PPF), on which noncommunists were represented, was a mere facade.

Improvements under Kádár

Nevertheless, conditions changed very much for the better. Kádár enunciated the principle that "he who is not against us is with us," which for the ordinary people meant that they could go about their business without fear of molestation or even much surveillance and could speak, read, and even write with reasonable freedom. Technical competence replaced party orthodoxy as a criterion for posts of responsibility. More scope was allowed to private small-scale enterprise in trade and industry, and the New Economic Mechanism (NEM), initiated in 1968, even introduced the profit motive into state-directed enterprises. Agricultural cooperatives were allowed to produce industrial goods for their own use or to sell on demand, while the private plots of their members supplied a large proportion of fruits and vegetables for the rest of the population. Contacts with the West were encouraged. A modus vivendi was found with the Vatican and with Protestant churches. The standard of living began to rise substantially. Tourism developed as a substantial industry. In 1988, for example, nearly 18 million foreigners visited Hungary, six million of them from western Europe, the United States, and Canada; and 10.8 million Hungarians traveled abroad, including 3.2 million to the West. With the introduction on Jan. 1, 1988, of "global passports," which removed restrictions on travel, the number of Hungarian tourists to the West more than trebled. In the same year income from tourism reached its highest level, yet the net balance was only about $100 million, since Hungarians going to the West spent most of their official hard currency quotas on purchases of consumer goods, due to shortages and skyrocketing prices at home.

The two decades of the NEM, which went beyond the liberalization that took place in the Soviet Union itself, were only partially successful. Productivity failed to rise according to expectations. Government regulations persisted in many areas, and the economy remained geared to the Soviet-led Comecon. A burdensome system of subventions aimed at keeping down the prices of basic necessities and services and at promoting the production of state-preferred goods made realistic cost accounting impossible. The price rise of oil and other industrial raw materials on the world market also aggravated the situation. The gap grew between the price of energy, sophisticated industrial hardware, and raw materials on the one hand and the price of agricultural products, a main item in Hungary's foreign trade, on the other. (C.A.M./G.Ba.)

Political opposition to reform, including Soviet and Comecon criticism of the NEM, all but brought it to a halt in 1973–78, when administrative interventions by state agencies and party and trade union organizations caused a return to the methods of the centralized command economy under the pretext of protecting the relative earnings of industrial workers compared to those in agriculture or of taxing only "unearned" profits of successful enterprises. Rezső Nyers, the architect of the NEM, was demoted in 1974, to be brought back to the Politburo only in May 1988, at a time of deepening political and economic crisis. By the end of the 1970s, reformers again prevailed over their opponents. New measures included cuts in the central bureaucracy, encouragement of small firms and private enterprises, revisions of the price and wage system to reflect more closely conditions on the world market and costs of production, and the creation of a commercial banking system.

Reforms of the late 1980s. *Economic reforms.* The efforts to introduce market reforms into Hungary's socialist economy extended to the international arena. Already a member of the General Agreement on Tariffs and Trade (GATT), Hungary was admitted to the International Monetary Fund (IMF) in 1982 and received assistance from the World Bank. An agreement with the European Economic Community (EEC), the first among members of Comecon, was also concluded. While the Soviet Union continued to be Hungary's single most important trading partner and the source of its energy supply, Hungary had to turn to the West for technological assistance and capital investment in the process of modernizing the economy. Trade relations with the West, in which Austria and West Germany played particularly important roles, were crucial at a time when barely half of Hungary's foreign trade involved members of Comecon. Foreign trade constituted a larger proportion of Hungary's gross national product (GNP) than that of any other Comecon country.

Economic deterioration

Efforts to adjust Hungary's economy to the world market were handicapped by the adverse effects of the energy crisis of the 1970s and the de facto reversal of the NEM in the same decade. Although agricultural production continued to advance, in part because of favourable international market conditions, the rest of the economy deteriorated. This process was further aggravated by misallocation of funds, reluctance to abandon costly projects such as the Danube hydroelectric power plant, participation in joint projects of Comecon, and unwillingness to drastically reduce subsidies to inefficient enterprises and for many basic necessities and services, whose price level was kept artificially low for many years. As a result Hungary's hard

currency debt reached $18 billion by the end of the 1980s, representing the highest per capita indebtedness of any country in eastern Europe. Inflationary pressures began to build up, and real wages and living standards declined.

The appointment of Károly Grósz as prime minister in mid-1987 led to a program of severe belt-tightening; a harsh, hastily prepared income-tax law aimed at cutting consumption; anticipated unemployment in some segments of the economy; and steep rises in consumer prices, transportation costs, and basic services such as gas, electricity, telephone, water, and rents. Minor changes in the party leadership, still controlled by Kádár, and the reshuffling of the government—including the establishment of the first Ministry of the Environment in eastern Europe—eased acceptance of unpopular measures introduced to stabilize the collapsing economy.

Following the establishment of a multiparty state in 1989–90, aid and investment from the European Communities, the United States, and Japan, together with a standby credit arrangement from the International Monetary Fund, provided some economic relief. A series of price increases aimed at deficit reduction was also implemented in 1990. The major obstacle facing the economy, however, was the need to direct a large percentage of trade away from the Comecon countries and toward markets in the West.

Political changes and reforms. By the late 1980s, growing numbers of Hungarians had concluded that years of misgovernment could not be erased by economic reforms alone. The process of de-Stalinization reinforced the desire to reexamine the political premises of Grósz's program, which seemed to imply that, to keep their hard-fought-for personal freedoms, Hungarians should pay with economic misery and further social polarization. By the time the rate of inflation reached 17 percent, public pressure compelled the party conference in May 1988 not only to replace Kádár with Grósz but also to replace several of Kádár's supporters in the Politburo and Central Committee (CC) with Nyers and Imre Pozsgay, the secretary of the PPF, as ministers in charge of the economic and sociocultural spheres, respectively. They kept their positions even after Miklós Németh took over the prime ministership from Grósz in November 1988. In 1989 a quadrumvirate of Grósz, Németh, Pozsgay, and Nyers, chaired by the latter, temporarily took over the direction of a deeply split party. In October the party congress announced its transformation into the Hungarian Socialist Party (HSP). A splinter group of conservatives retained the old name and continued allegiance to its policies.

Weakening of the Politburo

Meanwhile, informal associations and clubs, debating circles such as the Hungarian Democratic Forum, the Alliance of Young Democrats, the Network of Free Initiatives, and the Bajcsy-Zsilinszky Society proliferated and served as points of departure for new political parties. The Democratic Union of Scientific Workers, supported by a substantial portion of academic and clerical employees of scholarly institutions, was the first independent professional association to challenge the communist-controlled National Council of Trade Unions and to establish contact with the Polish Solidarity as well as organized labour in the West. Filmmakers, writers, and journalists rediscovered their right of free speech, publishers printed manuscripts kept locked up for decades, new periodicals appeared, and the press, radio, and television threw over taboos that had prevailed for more than 40 years.

The 950th anniversary of the death of Stephen I was celebrated with medieval pomp in August 1988 in the presence of the apostolic delegate of Pope John Paul II and the primate of Poland, Józef Cardinal Glemp. The PPF daily newspaper published a commemorative article by the Hungarian primate, László Cardinal Paskai, and the venerated relic of King Stephen, the Holy Dexter, was displayed in major towns. A long-standing demand of the Roman Catholic hierarchy was met by recognizing the special status of conscientious objectors in the armed forces. Full diplomatic relations with the Vatican were reestablished in 1990. The state returned to the Protestant churches some of their former prestigious educational institutions, and the Boy Scouts, a conspicuously Christian organization in Hungary, was resuscitated. A memorial honouring Raoul Wallenberg, the Swedish diplomat who rescued Jews in Nazi-occupied Hungary, was erected in 1986. The World Jewish Congress held its executive session in Budapest in 1987. Anti-Semitism, past and present, was discussed, and the same held true for the emotionally loaded social and racial question of the Gypsy population.

Officially, manifestations of ethnocentric nationalism were forbidden after 1945: a group of racists received prison sentences for insulting Cubans and seeking to rid Hungary of certain minorities, including Arabs. The Kádár regime had tried to avoid offending fraternal communist governments by not raising questions about violations of the rights of Hungarian minorities within their jurisdiction. When the writer Gyula Illyés defended the human rights of Hungarians in Transylvania during the 1970s, he was attacked in the party press. But, in the late 1980s, Hungarian party and government leaders joined public protests against this repression and the implementation of Romania's policy of reapportionment and relocation of the rural population, which affected a disproportionately large number of ethnic Hungarian settlements. By granting asylum to refugees from Transylvania (including Romanians and Germans) at a moment of economic insecurity, by tolerating if not encouraging a sharp media campaign and mass demonstrations in front of the Romanian Embassy in Budapest, and by submitting formal complaints to international organizations after an unsuccessful meeting between Grósz and Nicolae Ceauşescu in the Romanian city of Arad in August 1988, the Hungarian government indicated its determination to take an active interest in the fate of Hungarian minorities in neighbouring countries. The rift between the two countries deepened as ethnic disturbances in Transylvania continued even after the fall of the Ceauşescu regime. Hungary declared all Romanians of Hungarian descent to be dual citizens, in defiance of a 1979 bilateral agreement. This policy in defense of human rights, combined with renewed openings toward Austria, establishment of trade relations with South Korea, and the resumption of diplomatic relations with Israel (severed since the Arab-Israeli War of 1967), was taken as a sign of a more independent foreign policy, as were the efforts at strengthening Hungary's ties with the EEC.

Repression of Hungarians in Transylvania

In 1989 the Hungarian section of the Iron Curtain along the Austrian border was dismantled. In July, George Bush, the first U.S. president to set foot on Hungarian soil, was welcomed in Budapest. Interviews broadcast by Hungarian television with Alexander Dubček and Ota Šik, leaders of the Czechoslovak reform movement, and with the exiled king Michael of Romania, as well as Hungary's refusal to prevent tens of thousands of East German refugees from crossing into Austria on their way to West Germany, led to formal protests in Prague, Bucharest, and East Berlin.

The changes on the domestic scene were no less dramatic. They extended to the constitutional framework built since the communist takeover. Guidelines for a new constitution, drafted by the government and approved by both the party and the National Assembly, did not mention the "leading role of the Party," spelled out by the constitution of 1949. The draft of the new constitution sanctioned a multiparty system that had already been accepted in principle by the party leadership. The new constitution was based on the separation of legislative, executive, and judicial powers and included guarantees of individual and civil rights. Many of these changes were carried into effect in October 1989 by a series of amendments to the 1949 constitution; other changes included the creation of the post of state president in place of the Presidential Council and the elimination of the word *People's* from the name of the country.

Constitutional amendments of 1989

Important new legislation included amendments to the law of assembly, which granted the holding of indoor meetings without special permission, and a new enterprise law, which allowed the private ownership of businesses with up to 500 employees, permitted foreigners to own up to 100 percent of an enterprise, and allowed mixed (*i.e.,* joint state and private) ownership of property. The government consulted with independent organizations and spokesmen of the opposition in the course of preparing the new laws,

and the function of the National Assembly underwent a certain change. Hitherto a rubber stamp, the Assembly rejected the government's budget during its autumn 1988 session, and a substantial part of its membership vigorously objected to further funding of the Gabcikovo–Nagymaros hydroelectric project, which had been severely criticized by environmental groups and others.

Alternative independent parties and organizations continued to grow. The Independent Smallholders' Party, the Social Democrats, the People's Party, and the Hungarian Independence Party, destroyed or emasculated by the communists in 1949, reemerged. A Christian Democratic Party also began to organize. Few of the new parties chose to support socialist principles. The same held true for the Young Hungarian Workers' Organization, the Federation of Agrarian Youth, the Alliance of Free Democrats, and the Hungarian Democratic Forum, whose membership passed 10,000 by the end of 1988.

The election of 1990

The focus of the new parties was on the 1990 general election, the first major free election in Hungary in more than 40 years. No single party captured a majority of seats, but a coalition of the Forum, Smallholders, and Christian Democrats was able to form a government under Forum leader József Antall. An accord between the Forum and the Free Democrats, the two leading parties, resulted in the selection of Free Democrat Árpád Göncz as acting state president. The two parties shared many positions but differed principally in the schedule with which they planned to implement economic and political reforms, the centre-right Forum favouring a lengthier transition. The Forum's appeal to nationalism, or "Hungarianness," was also a factor in its success. Despite its role in constitutional reforms, the HSP did poorly in the election, winning fewer than 10 percent of the seats in the National Assembly.

One of the promising symbolic victories won by the reformers, led by a Commission for Historical Justice, was the reburial of Nagy and five of his associates and the reevaluation of the revolution of 1956, along with the entire post-1945 history of Hungary. Indeed, tens of thousands of cases involving innocent persecuted persons are undergoing judicial review. The monuments to be erected in honour of the victims of 1956 were important, as were the questions raised in public for the first time about Kádár's role in the political murders of Rajk and Nagy. To overcome the mistakes and crimes of the past and to avoid their repetition in the future, Hungarians, like other nations, turned to the study of their history. On March 15, 1989, the anniversary of the anti-Habsburg revolution of 1848, commemorated as a national holiday for the first time in 40 years, there were two celebrations in many places, including Budapest: one was organized by officials who until recently had opposed such public celebrations of a truly national holiday, and one by demonstrators who refused to march together with those held responsible for past police brutalities.

Hungarians continued to be a divided nation, not only because one-third of them lived abroad. Nor were the divisions splitting them apart exclusively political. According to official data, one-fifth of Hungarian citizens lived in poverty; restoration of a multiparty system and even a free and democratic society would not bring relief overnight. Given the global crisis of modern communism, particularly acute in eastern Europe, Hungarians would need a good deal of courage and luck—*Sors bona* in the words of the 17th century poet-statesman Miklós Zrínyi—to overcome their problems and to save their self-respect.

(G.Ba./Ed.)

For later developments in the history of Hungary, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 921, 923, 961, 962, 963, and 972, and the *Index.*

BIBLIOGRAPHY

Physical and human geography. Overviews of the history, geography, and people and of the social, economic, and cultural life of the country are provided by EUGENE K. KEEFE *et al., Area Handbook for Hungary* (1973), still useful; STEPHEN BORSODY (ed.), *The Hungarians: A Divided Nation* (1988); *Hungary* (annual); and the *Statistical Pocket Book of Hungary* (annual). MÁRTON PÉCSI and BÉLA SÁRFALVI, *The Geography of Hungary* (1964; originally published in Hungarian, 1960), gives a detailed picture of the physical and economic geography. MÁRTON PÉCSI, *Geomorphological Regions of Hungary,* trans. from Hungarian (1970), is a brief description of individual regions. Economic studies include JOHN KOMLOS (ed.), *Economic Development in the Habsburg Monarchy in the Nineteenth Century: Essays* (1983); IVAN T. BEREND and GYÖRGY RÁNKI, *Hungary: A Century of Economic Development* (1974), and *The Hungarian Economy in the Twentieth Century* (1985); and TIVADAR BERNÁT (ed.), *An Economic Geography of Hungary,* trans. from Hungarian (1985). Hungary's cinematic achievements are detailed in GRAHAM PETRIE, *History Must Answer to Man: The Contemporary Hungarian Cinema* (1978).

History. *General:* There are three major Hungarian histories: SÁNDOR SZILÁGYI (ed.), *A Magyar nemzet története,* 10 vol. (1895–98); BÁLINT HÓMAN and GYULA SZEKFŰ, *Magyar történet,* 7th ed., 5 vol. (1941–43); and *Magyarország története* (1976–), of which 5 vol. have been published, covering the period 1848–1945. Major sections dealing with Hungary appear in ADAM WANDRUSZKA and PETER URBANITSCH (eds.), *Die Habsburgermonarchie 1848–1918* (1973–). Histories in English include DOMOKOS G. KOSÁRY, *A History of Hungary* (1941, reissued 1971); DENIS SINOR, *History of Hungary* (1959, reprinted 1976); C.A. MACARTNEY, *Hungary, a Short History* (1962); PAUL IGNOTUS, *Hungary* (1972); and JÖRG HOENSCH, *A History of Modern Hungary, 1867–1986* (1988; originally published in German, 1984).

Early history: C.A. MACARTNEY, *The Magyars in the Ninth Century* (1930, reprinted 1968). *18th and 19th centuries:* HENRY MARCZALI, *Hungary in the Eighteenth Century* (1910, reprinted 1971; originally published in Hungarian, 1882); DENIS SILAGI, *Ungarn und der geheime Mitarbeiterkreis Kaiser Leopolds II.* (1961); DOMOKOS G. KOSÁRY, *Culture and Society in Eighteenth Century Hungary* (1987); C.A. MACARTNEY, *The Hapsburg Empire, 1790–1918* (1968); PETER HANÁK, *Ungarn in der Donaumonarchie: Probleme der bürgerlichen Umgestaltung einer Vielvölkerstaates* (1984); GEORGE BARANY, *Stephen Széchenyi and the Awakening of Hungarian Nationalism, 1791–1841* (1968); GYÖRGY SPIRA, *A Hungarian Count in the Revolution of 1848,* trans. from Hungarian (1974); PAUL BÖDY, *Joseph Eötvös and the Modernization of Hungary 1840–1870: A Study of Ideas of Individuality and Social Pluralism in Modern Politics,* 2nd rev. ed. (1985); BELA K. KIRALY, *Ferenc Deák* (1975); ISTVÁN DEÁK, *The Lawful Revolution: Louis Kossuth and the Hungarians, 1848–1849* (1979); GABOR VERMES, *István Tisza: The Liberal Vision and Conservative Statecraft of a Magyar Nationalist* (1985); R.W. SETON-WATSON, *Racial Problems in Hungary* (1908, reprinted 1972). *World War I and the Treaty of Trianon:* MICHAEL KÁROLYI, *Fighting the World: The Struggle for Peace,* trans. from Hungarian (1923); RUDOLF L. TÖKÉS, *Béla Kun and the Hungarian Soviet Republic: The Origins and Role of the Communist Party of Hungary in the Revolutions of 1918–1919* (1967); FRANCIS DEÁK, *Hungary at the Paris Peace Conference: The Diplomatic History of the Treaty of Trianon* (1942, reissued 1972); C.A. MACARTNEY, *Hungary and Her Successors: The Treaty of Trianon and Its Consequences, 1919–1937* (1937, reprinted 1968). *1920–45:* C.A. MACARTNEY, *October Fifteenth: A History of Modern Hungary, 1929–1945,* 2nd ed., 2 vol. (1961; also published as *A History of Hungary, 1929–1945,* 2 vol., 1956–57); THOMAS L. SAKMYSTER, *Hungary, the Great Powers, and the Danubian Crisis, 1936–1939* (1980); MARIO D. FENYO, *Hitler, Horthy, and Hungary: German-Hungarian Relations, 1941–1944* (1972); GYULA JUHÁSZ, *Hungarian Foreign Policy, 1919–1945* (1979); RANDOLPH L. BRAHAM, *The Politics of Genocide: The Holocaust in Hungary,* 2 vol. (1981); ANDREW C. JANOS, *The Politics of Backwardness in Hungary, 1825–1945* (1982). *Hungary under communism:* ERNST C. HELMREICH (ed.), *Hungary* (1957, reissued 1973); MIKLÓS MOLNÁR, *A Short History of the Hungarian Communist Party* (1978); BENNETT KOVRIG, *Communism in Hungary: From Kun to Kádár* (1979). *Revolution of 1956:* FERENC A. VÁLI, *Rift and Revolt in Hungary: Nationalism Versus Communism* (1961); PAUL E. ZINNER, *Revolution in Hungary* (1962, reissued 1972); CHARLES GATI, *Hungary and the Soviet Bloc* (1986); WILLIAM F. ROBINSON, *The Pattern of Reform in Hungary: A Political, Economic, and Cultural Analysis* (1973); PETER A. TOMA and IVAN VOLGYES, *Politics in Hungary* (1977). Further references may be found in ELEMER BAKO, *Guide to Hungarian Studies,* 2 vol. (1973); and STEVEN BELA VARDY, *Modern Hungarian Historiography* (1976).

(C.A.M./G.Ba.)

The Hydrosphere

Water is the most abundant substance at the surface of the Earth. Approximately one billion cubic kilometres (336 million cubic miles) of water in liquid and frozen form make up the oceans, lakes, streams, glaciers, and groundwaters found there. It is this enormous volume of water, in its various manifestations, that forms the discontinuous layer, enclosing much of the terrestrial surface, known as the hydrosphere.

Central to any discussion of the hydrosphere is the concept of the hydrologic cycle. This cycle consists of a group of reservoirs containing water, the processes by which water is transferred from one reservoir to another (or transformed from one state to another), and the rates of transfer associated with such processes. These transfer paths penetrate the entire hydrosphere, extending upward to about 15 kilometres (nine miles) in the Earth's atmosphere and downward to depths on the order of five kilometres in its crust.

This article examines the processes of the hydrologic cycle and discusses the way in which the various reservoirs of the hydrosphere are related through the hydrologic cycle. It also describes the biogeochemical properties of the waters of the Earth at some length and considers the distribution of global water resources and their utilization and pollution by human society. Details concerning the major water environments that make up the hydrosphere are provided in the articles OCEANS; LAKES; RIVERS; and ICE AND ICE FORMATIONS. See also CLIMATE AND WEATHER for specific information about the impact of climatic factors on the hydrologic cycle. The principal concerns and methods of hydrology and its various allied disciplines are summarized in EARTH SCIENCES. For coverage of other related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 221, 222, 223, 351, and 10/33, and the *Index.*

The article is divided into the following sections:

Distribution and quantity of the Earth's waters 715
Biogeochemical properties of the hydrosphere 715
  Rainwater
  River and ocean waters
  Lake waters
  Groundwaters
  Ice
The hydrologic cycle 720
  General nature of the cycle
  Processes involved in the cycle
Origin and evolution of the hydrosphere 725
  The early hydrosphere
  The transitional hydrosphere
  The modern hydrosphere
Impact of human activities on the hydrosphere 728
  Eutrophication
  Acid rain
  Buildup of greenhouse gases
Bibliography 731

DISTRIBUTION AND QUANTITY OF THE EARTH'S WATERS

Ocean waters and waters trapped in the pore spaces of sediments make up most of the present-day hydrosphere (see Table 1). The total mass of water in the oceans equals about 50 percent of the mass of sedimentary rocks now in existence and about 5 percent of the mass of the Earth's crust as a whole. Deep and shallow groundwaters constitute a small percentage of the total water locked in the pores of sedimentary rocks—on the order of 3 to 15 percent. The amount of water in the atmosphere at any one time is trivial, equivalent to 0.013×10^6 cubic kilometres of liquid water, or about 0.001 percent of the total at the Earth's surface. This water, however, plays an important role in the water cycle.

At present, ice locks up a little more than 1 percent of the Earth's water and may have accounted for as much as 3 percent or more during the height of the glaciations of the Pleistocene epoch (from 1,600,000 to 10,000 years ago). Although water storage in rivers, lakes, and the atmosphere is small, the rate of water circulation through the rain–river–ocean–atmosphere system is relatively rapid. The amount of water discharged each year into the oceans from the land is approximately equal to the total mass of water stored at any instant in rivers and lakes.

Soil moisture accounts for only 0.005 percent of the water at the Earth's surface. It is this small amount of water, however, that exerts the most direct influence on evaporation from soils. The biosphere, though primarily H_2O in composition, contains very little of the total water at the terrestrial surface, only about 0.00004 percent. Yet, the biosphere plays a major role in the transport of water vapour back into the atmosphere by the process of transpiration.

As will be seen in the next section, the Earth's waters are not pure H_2O but contain dissolved and particulate materials. Thus, the masses of water at the Earth's surface are major receptacles of inorganic and organic substances, and water movement plays a dominant role in the transportation of these substances about the planet's surface.

Table 1: Water Masses at the Earth's Surface

reservoir	volume (in millions of cubic kilometres)	percent of total
Oceans	1,370	97.25
Ice caps and glaciers	29	2.05
Deep groundwater* (750–4,000 metres)	5.3	0.38
Shallow groundwater (< 750 metres)	4.2	0.30
Lakes	0.125	0.01
Soil moisture	0.065	0.005
Atmosphere†	0.013	0.001
Rivers	0.0017	0.0001
Biosphere	0.0006	0.00004
Total	1,408.7	100

*The total interstitial water in the pores of sediments is on the order of $50–300 \times 10^6$ km^3. †As liquid equivalent of water vapour.
Source: Adapted from Elizabeth Kay Berner and Robert A. Berner, *The Global Water Cycle: Geochemistry and Environment* (1987). Prentice Hall, Inc.

BIOGEOCHEMICAL PROPERTIES OF THE HYDROSPHERE

Rainwater. About 110,300 cubic kilometres of rain fall on land each year. The total water in the atmosphere is 0.013×10^6 cubic kilometres, and this water, due to precipitation and evaporation, turns over every 9.6 days. Rainwater is not pure but rather contains dissolved gases and salts, fine-ground particulate material, organic substances, and even bacteria. The sources of the materials in rainwater are the oceans, soils, fertilizers, air pollution, and fossil-fuel combustion.

Constituents of rainwater

Figure 1 shows some typical maps of the chloride and sodium content of rain over the continental United States. The concentration contours in Figure 1 are subparallel to the coastlines. It also has been observed that rains over oceanic islands and near coasts have ratios of major dissolved constituents very close to those found in seawa-

ter. The discovery of the high salt content of rain near coastlines was somewhat surprising because sea salts are not volatile, and it might be expected that the process of evaporation of water from the sea surface would "filter" out the salts. It has been demonstrated, however, that a large percentage of the salts in rain is derived from the bursting of small bubbles at the sea surface due to the impact of rain droplets or the breaking of waves, which results in the injection of sea aerosol into the atmosphere. This sea aerosol evaporates, with resultant precipitation of the salts as tiny particles that are subsequently carried high into the atmosphere by turbulent winds. These particles may then be transported over continents to fall in rain or as dry deposition.

Assuming equilibrium with the atmospheric carbon dioxide partial pressure (P_{CO_2}) of $10^{-3.5}$ atmosphere, the approximate mean composition of rainwater is in parts per million (ppm): sodium (Na^+), 1.98; potassium (K^+), 0.30; magnesium (Mg^{2+}), 0.27; calcium (Ca^{2+}), 0.09; chloride (Cl^-), 3.79; sulfate (SO_4^{2-}), 0.58; and bicarbonate (HCO_3^-), 0.12. In addition to these ions, rainwater contains small amounts of dissolved silica—about 0.30 ppm. The average pH value of rainwater is 5.7. (The term pH is defined as the negative logarithm of the hydrogen ion concentration in moles per litre. The pH scale ranges from 0 to 14, with lower numbers indicating increased acidity.) On a global basis, as much as 35 percent of the sodium, 55 percent of the chlorine, 15 percent of the potassium, and 37 percent of the sulfate in river water may be derived from the oceans through sea aerosol generation.

ignificance f marine erosols

A considerable amount of data has become available for marine aerosols. These aerosols are important because (1) they are vital to any description of the global biogeochemical cycle of an element; (2) they may have an impact on climate; (3) they are a sink, via heterogeneous chemical reactions, for trace atmospheric gases; and (4) they influence precipitation of cloud and rain droplets. For many trace metals the ratio of the atmospheric flux to the riverine flux for coastal and remote oceanic areas may be greater than one, indicating the importance of atmospheric transport. Figure 2 illustrates the enrichment factors (EF) of North Atlantic marine aerosols and suspended matter in North Atlantic waters relative to the crust, where

$$\mathrm{EF}_{\mathrm{crust}} = \frac{(X/\mathrm{Al})_{\mathrm{air}}}{(X/\mathrm{Al})_{\mathrm{crust}}},$$

and $(X/\mathrm{Al})_{\mathrm{air}}$ and $(X/\mathrm{Al})_{\mathrm{crust}}$ refer, respectively, to the ratio of the concentration of the element X to that of Al, aluminum, in the atmosphere and in average crustal material. The similarity in trend of enrichment factors for marine aerosols and suspended matter indicates qualitatively the importance of the marine aerosol to the composition of marine suspended matter and, consequently, to deep-sea sedimentation.

From P. Buat-Menard, "Particle Geochemistry in the Atmosphere and Oceans," *Air-Sea Exchange of Gases and Particles* (1983); D. Reidel Publishing Co.

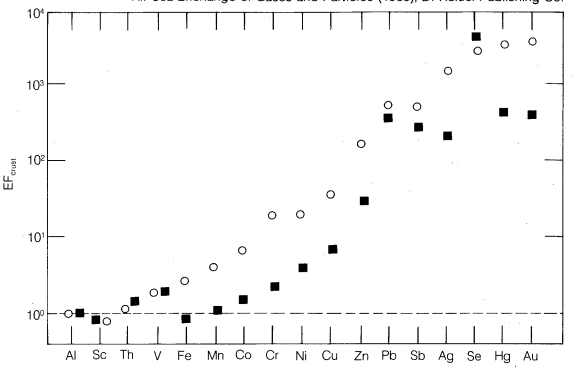


Figure 2: Element enrichment factors (EF) in North Atlantic marine aerosols (squares) compared with factors for suspended matter in North Atlantic ocean waters (circles).

In some instances the ratios of ions in rainwater deviate significantly from those in seawater. Mechanisms proposed for this fractionation are, for example, the escape of chlorine as gaseous hydrogen chloride (HCl) from sea-salt aerosol with a consequent enrichment in sodium and bubbling and thermal diffusion. In addition, release of gases like dimethyl sulfide (DMS) from the sea surface and its subsequent reaction in the oceanic atmosphere to sulfate can change rainwater ion ratios with respect to seawater. Soil particles also can influence rainwater composition. Rainfall over the southwestern United States contains relatively high sulfate concentrations because of sulfate-bearing particles that have been blown into the atmosphere from desert soils. Rain near industrial areas commonly contains high contents of sulfate, nitrate, and carbon dioxide (CO_2) largely derived from the burning of coal and oil. There are two main processes leading to the conversion of sulfur dioxide (SO_2) to sulfuric acid (H_2SO_4). These are reactions with hydroxyl radicals (OH) and with hydrogen peroxide (H_2O_2) in the atmosphere:

$$SO_2 + OH \rightarrow \text{intermediate species} \rightarrow H_2SO_4 \qquad (1)$$

and

$$SO_2 + H_2O_2 = H_2SO_4. \qquad (2)$$

The sulfuric acid then dissociates to hydrogen and sulfate ions:

$$H_2SO_4 = 2H^+ + SO_4^{2-}. \qquad (3)$$

For the nitrogen gases nitric oxide (NO) and nitrogen dioxide (NO_2) released from fossil-fuel burning, their atmospheric reactions lead to the production of nitric acid (HNO_3) and its dissociation to hydrogen ions (H^+) and nitrate (NO_3^-). These reactions are responsible for the acid rain conditions highly evident in the northeastern United States, southeastern Canada, and western Europe (see below *Acid rain*). The high sulfate values of the rain in the northeastern United States reflect the acid precipitation conditions of this region.

Adapted from data from the National Atmospheric Deposition Program (NADP)

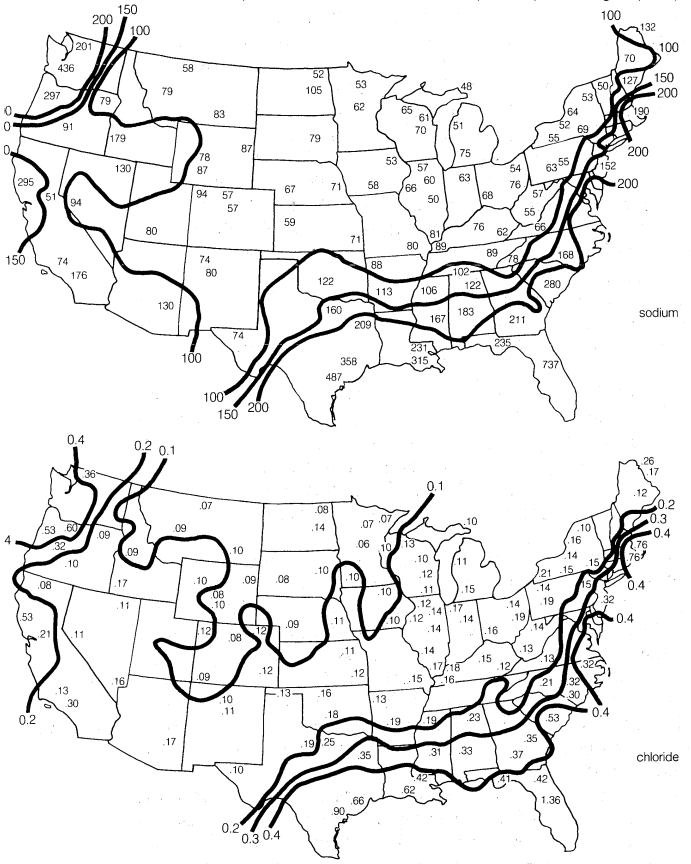


Figure 1: Contours of the average concentrations of dissolved sodium and chloride in precipitation over the United States during 1987.

River and ocean waters. River discharge constitutes the main source for the oceans. Table 2 shows the difference between the composition of seawater and that of river, or

Difference in composition

stream, water. Seawater has a more uniform composition than river water. It contains, by weight, about 3.5 percent dissolved salts, whereas river water has only 0.012 percent. The average density of the world's oceans is roughly 2.75 percent greater than that of typical river water. Of the average 35 parts per thousand salts of seawater, sodium and chlorine make up almost 30 parts, and magnesium and sulfate contribute another four parts. Of the remaining one part of the salinity, calcium and potassium constitute 0.4 part each and carbon, as carbonate and bicarbonate, about 0.15 part. Thus, only eight elements (oxygen, sulfur, chlorine, sodium, magnesium, calcium, potassium, and carbon) make up 99 percent of seawater, though most of the 92 naturally occurring elements have been detected therein. Of importance are the nutrient elements phosphorus, nitrogen, and silicon, along with such essential micronutrient trace elements as iron, cobalt, and copper. These elements strongly regulate the organic production of the world's oceans.

In contrast to ocean water, the average salinity of the world's rivers is low—only about 0.012 percent, or 120 ppm by weight. Of this salt content, carbon as bicarbonate constitutes 58 parts, or 48 percent, and calcium, sulfur as sulfate, and silicon as dissolved monomeric silicic acid make up a total of about 39 parts, or 33 percent. The remaining 19 percent consists predominantly of chlorine, sodium, and magnesium in descending importance. It is obvious that the concentrations and relative proportions of dissolved species in river waters contrast sharply with those of seawater. Thus, even though seawater is derived in part by the chemical differentiation and evaporation of river water, the processes involved affect every element differently, indicating that simple evaporation and concentration are entirely secondary to other processes.

Generally speaking, the composition of river water, and thus that of lakes, is controlled by water–rock interactions. The attack of carbon dioxide-charged rain and soil waters on the individual minerals in continental rocks leads to the production of dissolved constituents for lakes, rivers, and streams. It also gives rise to solid alteration products that make up soils or suspended particles in freshwater aquatic systems. The carbon dioxide content of rain and soil waters is of particular importance in weathering processes. The pH of rainwater equilibrated with the atmospheric carbon dioxide partial pressure of $10^{-3.5}$ atmosphere is 5.7. In industrial regions, rainwater pH values may be lower because of the release and subsequent hydrolysis of acid gases—namely, sulfur dioxide and nitrogen oxides (NO_x) from the combustion of fossil fuels. After rainwater enters soils, its characteristics change markedly. The usual few parts per million of salts in rainwater increase substantially as the water reacts. The upper part of the soil is a zone of intense biochemical activity. The bacterial population near the surface is large, but it decreases rapidly downward with a steep gradient. One of the major biochemical processes of the bacteria is the oxidation of organic material, which

Table 2: Composition of Oceans and Streams

atomic number	atomic weight	element	seawater (in micrograms per litre or parts per billion)	streams* (in micrograms per litre or parts per billion)
1	1.00797	hydrogen	1.1×10^8	1.10×10^8
2	4.0026	helium	0.0068	†
3	6.939	lithium	180	3
4	9.0122	beryllium	0.0056	†
5	10.811	boron	4,440	18
6	12.01115	carbon	28,000	11,620
		(inorganic) and dissolved organic	500	10,800
7	14.0067	nitrogen	15,000	950
		(dissolved N_2, NO_3^-, NO_2^-, NO_4^+), and dissolved organic	670	226
8	15.9994	oxygen	6,000	†
		(dissolved O_2, H_2O)	8.83×10^8	8.83×10^8
9	18.9984	fluorine	1,300	100
10	20.183	neon	0.120	†
11	22.9898	sodium	1.08×10^7	7,200
12	24.312	magnesium	1.29×10^6	3,650
13	26.9815	aluminum	1	50
14	28.086	silicon	2,900	4,850
15	30.9738	phosphorus	60	78
16	32.064	sulfur	9.05×10^5	3,830
17	35.453	chlorine	1.9×10^7	8,250
18	39.948	argon	430	†
19	39.102	potassium	3.8×10^5	1,400
20	40.08	calcium	4.12×10^5	14,700
21	44.956	scandium	0.0006	0.004
22	47.90	titanium	1	10
23	50.942	vanadium	2.5	0.9
24	51.996	chromium	0.3	1
25	54.9380	manganese	0.2	8.2
26	55.847	iron	2	40
27	58.9332	cobalt	0.05	0.2
28	58.71	nickel	1.7	2.2
29	63.54	copper	0.5	10
30	65.37	zinc	4.9	30
31	69.72	gallium	0.03	0.09
32	72.59	germanium	0.05	†
33	74.9216	arsenic	3.7	2
34	78.96	selenium	0.2	0.2
35	79.909	bromine	67,300	20
36	83.80	krypton	0.21	†
37	85.47	rubidium	120	1.5
38	87.62	strontium	8,100	60
39	88.905	yttrium	0.013	0.07
40	91.22	zirconium	0.026	†
41	92.906	niobium	0.015	†
42	95.94	molybdenum	10	0.5
43	98.906	technetium	(not naturally occurring)	
44	101.07	ruthenium	0.0007	†
45	102.905	rhodium	†	†
46	106.4	palladium	†	†
47	107.87	silver	0.04	0.3
48	112.4	cadmium	0.11	†
49	114.82	indium	0.0001	†
50	118.69	tin	0.01	†
51	121.75	antimony	0.24	1
52	127.60	tellurium	†	†
53	126.90	iodine	64	7
54	131.30	xenon	0.05	†
55	132.91	cesium	0.40	0.035
56	137.34	barium	15	60
57	138.91	lanthanum	0.0034	0.05
58	140.12	cerium	0.0012	0.08
59	140.91	praseodymium	0.00064	0.007
60	144.24	neodymium	0.0028	0.04
61	(147)	promethium	(not naturally occurring)	
62	150.35	samarium	0.00045	0.008
63	151.96	europium	0.000130	0.001
64	157.25	gadolinium	0.00070	0.008
65	158.92	terbium	0.00014	0.001
66	162.50	dysprosium	0.00091	0.05
67	164.93	holmium	0.00022	0.001
68	167.26	erbium	0.00087	0.004
69	168.93	thulium	0.00017	0.001
70	173.04	ytterbium	0.00082	0.004
71	174.97	lutetium	0.00015	0.001
72	178.49	hafnium	<0.007	†
73	180.95	tantalum	<0.0025	†
74	183.85	tungsten	<0.1	0.03
75	186.2	rhenium	0.0084	†
76	190.2	osmium	†	†
77	192.2	iridium	†	†
78	195.09	platinum	†	†
79	196.97	gold	0.004	0.002
80	200.59	mercury	0.03	0.07
81	204.37	thallium	0.01	†
82	207.19	lead	0.03	1
83	208.98	bismuth	0.02	†
84–89 and 91		(thorium and uranium decay series elements: polonium, astatine, radon, francium, radium, actinium, and protactinium)		
90	232.04	thorium	<0.01	0.1
92	238.03	uranium	3.2	0.04

*Not corrected for anthropogenic influence. †Little data or no reasonable estimates available.

Source: Modified from R.M. Garrels, F.T. Mackenzie, and C. Hunt, *Chemical Cycles and the Global Environment.* Copyright © 1975 by W. Kaufmann, Inc. All rights reserved.

leads to the release of carbon dioxide. Soil gases obtained above the zone of water saturation may contain 10 to 40 times as much carbon dioxide as the free atmosphere, and in some cases carbon dioxide has been shown to make up 30 percent of the soil gases as opposed to 0.03 percent of the free atmosphere. In addition to the acid effects of carbon dioxide, a highly acidic microenvironment is created by the roots of living plants. Values of pH as low as 2 have been measured immediately adjacent to root hairs. Plants may have several kilometres of root hairs, and so their chemical effects are formidable.

Congruent and incongruent weathering reactions

These acid solutions in the soil environment attack the rock minerals, the bases of the system, producing neutralization products of dissolved constituents and solid particles. Two general types of reaction occur: congruent and incongruent. In the former, a solid dissolves, adding elements to the water in their proportions in the mineral. An example of such a weathering reaction is the solution of calcite ($CaCO_3$) in limestones:

$$CaCO_3 + CO_2(g) + H_2O = Ca^{2+} + 2HCO_3^-. \quad (4)$$

Here, one of the HCO_3^- ions comes from calcite and the other from $CO_2(g)$ in the reacting water. The amount of carbon dioxide dissolved according to reaction (4) depends on temperature, pressure, original bicarbonate content of the weathering solution, and the partial pressure of the carbon dioxide. The carbon dioxide and the temperature are the most important variables. Increases in one or both of these variables lead to increases in the amount of calcite dissolved. For example, for a carbon dioxide pressure of $10^{-3.5}$ atmosphere, the amount of calcium that can be dissolved until saturation is about $10^{-3.3}$ mole, or 20 ppm, at 25° C (77° F). For an atmospheric carbon dioxide pressure of 10^{-2} atmosphere and for a soil atmosphere of nearly pure carbon dioxide, the values are 65 and 300 ppm, respectively. The weathering of calcite leads to the release of calcium and bicarbonate ions into soil waters and groundwaters, and these constituents eventually reach lake and river systems. The insoluble residue of quartz (SiO_2), clay minerals (cation, aluminum, silicon, oxygen, hydrogen phases), and iron oxides (*e.g.,* FeOOH) in the limestone rock make up the deep-red soils that form from limestone weathering. These particles may be carried into streams by runoff and hence to lakes and the oceans and become part of the suspended load of these systems.

An example of an incongruent weathering reaction is that involving aluminosilicates. One such reaction is the aggressive attack of carbon dioxide-charged soil water on the mineral K-spar ($KAlSi_3O_8$), an important phase found in continental rocks. The reaction is

$$2KAlSi_3O_8 + 2CO_2 + 11H_2O = Al_2Si_2O_5(OH)_4 + 2K^+ + 2HCO_3^- + 4H_4SiO_4. \quad (5)$$

It should be noted that the K-spar changes into a new mineral—kaolinite (a clay mineral) in this case—plus solution, and acid is consumed. The total dissolved material per litre of soil solution released is about 60 ppm for a solution initially equilibrated with a typical soil carbon dioxide content. The water resulting from reaction (5) would contain bicarbonate, potassium, and dissolved silica in the ratios 1:1:2, and the new solid, kaolinite, would be a weathering product. These dissolved constituents and the solid alteration product would eventually reach rivers to be transferred possibly to lakes and ultimately to the sea. It has been demonstrated that the composition of river water is the product of a variety of mineral–water reactions such as (4) and (5). The dissolved load of the world's rivers comes from the following sources: 7 percent from beds of halite (NaCl) and salt disseminated in rocks; 10 percent from gypsum ($CaSO_4 \cdot 2H_2O$) and anhydrite ($CaSO_4$) deposits and sulfate salts disseminated in rocks; 38 percent from limestones and dolomites; and 45 percent from the weathering of one silicate mineral to another. Of the bicarbonate ions in river water, 56 percent stems from the atmosphere, 35 percent from carbonate minerals, and 9 percent from the oxidative weathering of fossil organic matter. Reactions involving silicate minerals account for 30 percent of the riverine bicarbonate ions.

Besides dissolved substances, rivers also transport solids in traction (*i.e.,* bed load) and, most importantly, suspended load. The present global river-borne flux of solids to the oceans is estimated as 155×10^{14} grams per year. Most of this flux comes from Southeast Asian rivers. The composition of this suspended material resembles soils and shales and is dominated by silicon and aluminum. Present elemental fluxes are estimated in 10^{12} grams per year as silicon, 4,420; aluminum, 1,460; iron, 740; calcium, 330; potassium, 310; magnesium, 210; and sodium, 110. The total load of particulate organic carbon of the world's rivers is 180×10^{12} grams per year. The riverine fluxes of trace metals to the oceans are dominated by their occurrence in the particulate phase as opposed to the dissolved phase. The particulate matter in river water is an important source of silicon, aluminum, ion, titanium, rubidium, scandium, vanadium, the lanthanides, and other elements for deep-sea sediments.

Lake waters. Although lake waters constitute only a small percentage of the water in the hydrosphere, they are an important ephemeral storage reservoir for fresh water. Aside from their recreational use, lakes constitute a source of water for household, agricultural, and industrial uses. Lake waters are also very susceptible to changes in chemical composition due to these uses and to other factors.

In general, fresh waters at the continental surface evolve from their rock sources by enrichment in calcium and sodium and by depletion in magnesium and potassium. In very soft waters the alkalies may be more abundant than the alkaline earths, and in the more concentrated waters of open river systems $Ca > Mg > Na > K$. For the anions, in general, HCO_3^- exceeds SO_4^{2-}, which is greater in concentration than Cl^-. It is worthwhile at this stage to consider some major mechanisms that control global surface water composition. These mechanisms are atmospheric precipitation, rock reactions, and evaporation-precipitation. They are illustrated in Figure 3, which shows schematically the envelope of world surface water compo-

Mechanisms responsible for controlling surface water composition

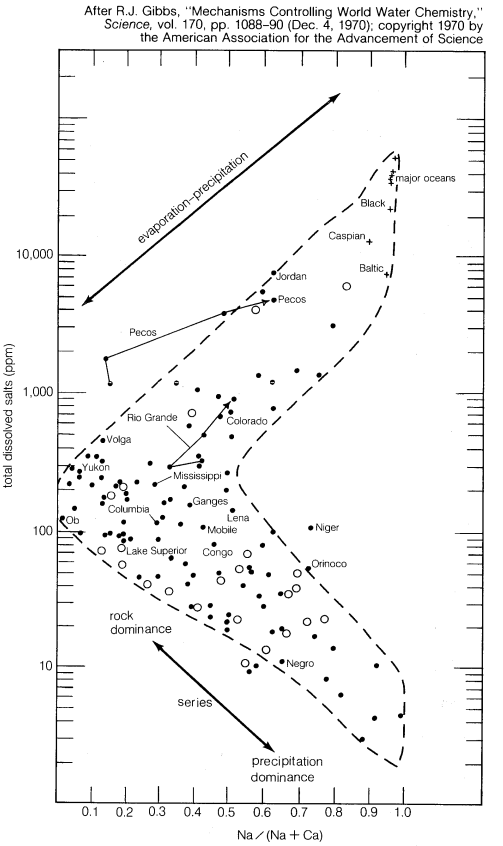


Figure 3: Variation in surface water composition as a function of total dissolved salt content and the ratio of dissolved sodium to sodium plus calcium in the waters. Three major mechanisms governing composition are shown in this diagram: evaporation-precipitation, precipitation, and rock–water reaction.

Table 3: Chemistry of Representative Closed-Basin, Saline Lake Waters*

	1 Carson Sink, Nevada	2 Bristol Lake, California	3 Salton Sea, California	4 Great Salt Lake, Utah	5 Surprise Valley Lake, California	6 Deep Springs Lake, California
SiO_2	19	—	20.8	48	36	—
Ca	261	43,296	505	241	11	3.1
Mg	129	1,061	581	7,200	31	1.2
Na	56,800	57,365	6,249	83,600	4,090	111,000
K	3,240	3,294	112	4,070	11	19,500
HCO_3	322	—	232	251	1,410	9,360
CO_3	—	—	—	—	664	22,000
SO_4	786	223	4,139	16,400	900	57,100
Cl	88,900	172,933	9,033	140,000	4,110	119,000
Total	152,000	279,150	20,900	254,000	10,600	335,000
pH	7.8			7.4	9.2	

*In parts per million.
Source: Modified from J.I. Drever, *The Geochemistry of Natural Waters*. © 1982. Prentice Hall, Inc.

sitions in terms of the variables of total dissolved salts and Na/(Na + Ca) ratio.

The mechanism principally responsible for waters of very low salinity is precipitation. These waters tend to form in tropical regions of low relief and thoroughly leached source rocks. In these regions rainfall is high, and water compositions are usually dominated by salts brought in by precipitation. Such waters constitute one end-member of a series of water compositions for which the other end-member represents water compositions dominated by contributions of dissolved salts from the rocks and soils of their basins. These waters have moderate salinity and are rich in dissolved calcium and bicarbonate. They are, in turn, the end-member of another series that extends from the calcium-rich, medium-salinity fresh waters to the high-salinity, sodium chloride-dominated waters of which seawater is an example. Seawater composition, however, does not evolve directly from the composition of fresh waters and the precipitation of calcium carbonate; other mechanisms that control its composition are involved. Such factors as relief and vegetation also may affect the composition of the world's surface waters, but atmospheric precipitation, water–rock reactions, and evaporation–crystallization processes appear to be the dominant mechanisms governing continental surface water chemistry.

Continental fresh waters evaporate once they have entered closed basins, and their constituent salts precipitate on the basin floors. The composition of these waters may evolve along several different paths, depending on their initial chemical makeup. Table 3 shows a number of brine compositions from North American saline lakes, and Figure 4 illustrates possible evaporation paths that led to these compositions. For example, Surprise Valley Lake, California, is a body of sodium chloride water that may have evolved from what was initially fresh water which precipitated calcite during evaporation, resulting in water with a dissolved-calcium concentration to alkalinity ratio

After L.A. Hardie and H.P. Eugster, "The Evolution of Closed-Basin Brines," *Fiftieth Anniversary Symposia: Mineralogy and Petrology of the Upper Mantle, Mineralogy and Geochemistry of Non-Marine Evaporites*, Special Paper No. 3, pp. 279 and 280, figs. 7 and 8, copyright by the Mineralogical Society of America, in J.I. Drever, *The Geochemistry of Natural Waters* (1982), Prentice Hall, Inc.

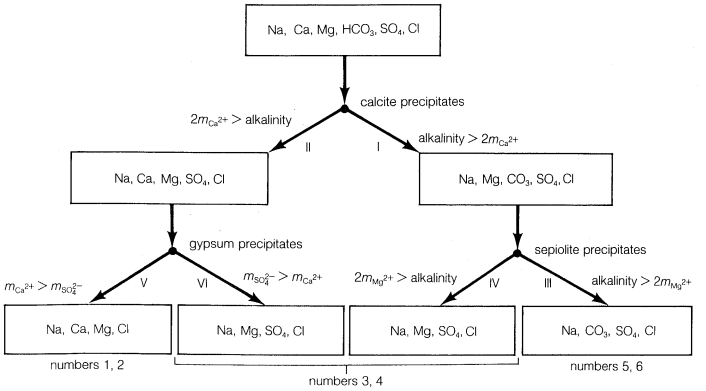


Figure 4: Possible evaporation paths leading to saline lake waters of different compositions. The numbers refer to the water compositions of Table 3.

of less than 2. Further evaporation of such water and precipitation of sepiolite [$MgSi_3O_6(OH)_2$] would give rise to water enriched in alkalinity with respect to dissolved magnesium and hence to sodium chloride–bicarbonate–carbonate water such as that of Surprise Valley Lake. The paths shown in Figure 4 are an oversimplification of the actual processes, which are far more diverse and complex in nature, but they do provide a reasonable picture of how fresh waters evolve into saline lake waters in closed basins.

Biological processes strongly affect the composition of lake waters and are responsible to a significant degree for the compositional differences between the upper water layer (the epilimnion) and the lower water layer (the hypolimnion) of lakes. The starting point is photosynthesis, represented by the following reaction:

$$106CO_2 + 16NO_3^- + HPO_4^{2-} + 122H_2O + 18H^+ \xrightarrow[\text{micro-nutrients}]{\text{radiant energy}} C_{106}H_{263}O_{110}N_{16}P + 138O_2. \quad (6)$$

The reversal of this reaction is oxidation-respiration leading to the release of the nutrients nitrogen and phosphorus, as well as carbon dioxide. In a stratified lake, carbon, nutrients, and silica are extracted from the epilimnion during photosynthesis. This process leads to reduced concentrations of nitrate, phosphate, and silica in these waters and, during times of maximum daylight organic production, supersaturation of epilimnion waters with respect to dissolved oxygen. The organic matter produced by phytoplankton may be either grazed upon by zooplankton and other organisms or decomposed by bacteria. Some of it, however, sinks into the hypolimnion. There, it is further decomposed, especially by bacteria, resulting in the release of dissolved phosphorus and nitrogen and the consumption of oxygen. Oxygen concentrations therefore are reduced in these lower lake waters, because stratification prevents oxygen exchange with the atmosphere. Furthermore, the inorganic carbonate and siliceous skeletons of the dead organisms sinking into the hypolimnion may dissolve, giving rise to increased concentrations of dissolved silica and inorganic carbon in the deep waters of stratified lakes. This dissolution is a result of undersaturation of the hypolimnion waters with respect to the opaline silica and calcium carbonate that make up the skeletons of the dead and sinking plankton. These natural biological processes have been accelerated in some lakes because of excess nutrient input by human activity, resulting in the eutrophication of lake waters (and marine systems; see below *Eutrophication*).

Processes affecting groundwater composition

Groundwaters. These waters derive their compositions from a variety of processes, including dissolution, hydrolysis, and precipitation reactions; adsorption and ion exchange; oxidation and reduction; gas exchange between groundwater and the atmosphere; and biological processes (see Table 4). The biological processes of greatest importance are microbial metabolism, organic production, and respiration (oxidation). By far the most important overall process for the major constituents of groundwater is that of mineral–water reactions, which were briefly described

above in *River and ocean waters.* Thus, the composition of groundwaters strongly reflects the types of rock minerals that the waters have encountered in their movement through the subsurface. Table 5 shows the waters found in limestones, crystalline rocks, and two types of fine-grained rocks. The mineralogy of the Wissahickon schist is dominated by aluminosilicate compositions, whereas the Ecca shale contains significant carbonate and dispersed salts. The latter minerals are more soluble than aluminosilicate minerals, and their dissolution gives rise to the high salinity of Ecca shale waters. The waters of the Wissahickon schist have low salinity partly because of the low chemical reactivity of the silicate minerals in this rock. In contrast, the Miocene limestone waters are dominated by dissolved calcium and bicarbonate, a characteristic reflecting the higher solubility and rate of dissolution of the calcite that makes up this rock. The groundwater from the granite is rich in the Ca^{2+} and Na^{+} cations derived from dissolution of the plagioclase feldspar (a sodium, calcium aluminosilicate), which is a major mineral found in this rock type. In general, the most mobile elements in groundwater—*i.e.,* those most easily liberated by the weathering of rock minerals—are calcium, sodium, and magnesium. Silicon and potassium have intermediate mobilities, and aluminum and iron are essentially immobile and locked up in solid phases.

Table 4: Processes Affecting the Major Chemical Components of Groundwater

component	origin*
Na^{+}	sodium chloride dissolution (some pollutive) plagioclase weathering rainwater addition
K^{+}	biotite weathering K-feldspar weathering biomass decreases dissolution of trapped aerosols
Mg^{2+}	amphibole and pyroxene weathering biotite (and chlorite) weathering dolomite weathering olivine weathering rainwater addition
Ca^{2+}	calcite weathering plagioclase weathering dolomite weathering dissolution of trapped aerosols biomass decreases
HCO_3^-	calcite and dolomite weathering silicate weathering
SO_4^{2-}	pyrite weathering (some pollutive) calcium sulfate dissolution rainwater addition
Cl^-	sodium chloride dissolution (some pollutive) rainwater addition
H_4SiO_4	silicate weathering

*The sources for each constituent are given in approximate order of decreasing importance.
Source: Adapted from Elizabeth Kay Berner and Robert A. Berner, *The Global Water Cycle: Geochemistry and Environment.* © 1987. Prentice Hall, Inc.

Groundwaters are highly susceptible to contamination because of human activities and the fact that their dissolved constituents are derived to a large extent from the leaching of surface materials. Some of the nitrogen and phosphorus applied to soils as fertilizers and organic pesticides may be leached and leak into groundwater systems, leading to increased concentrations of ammonium and phosphate. Radioactive wastes, industrial chemicals, household materials, and mine refuse are other anthropogenic sources of dissolved substances that have been detected in groundwater systems.

Ice. Ice is nearly a pure solid and, as such, accommodates few foreign ions in its structure. It does contain, however, particulate matter and gases, which are trapped in bubbles within the ice. The change in composition of these materials through time, as recorded in the successive layers of ice, has been used to interpret the history of the Earth's surface environment and the impact of human activities on this environment. The increase in the lead content of continental glacial ice with decreasing age of the ice, for example, reflects the progressive input of tetraethyl lead into the global environment from gasoline burning. Also, atmospheric carbon dioxide and methane concentrations, which have increased significantly during the past century because of anthropogenic activities, are faithfully recorded in ice bubbles of the thick continental ice sheets. Present-day atmospheric carbon dioxide and methane concentrations are 25 percent and 167 percent, respectively, higher than their concentrations 200 years ago; the latter concentration values were obtained from measurements of the gases in air trapped in ice.

Table 5: Groundwater Compositions in Major Rock Types

	limy	argillaceous	salty	crystalline
	Miocene limestone	Wissahickon schist	Ecca shale, South Carolina	granite, South Carolina
SiO_2	8.9	14.0	32	35.0
Ca^{2+}	48.0	3.1	62	13.0
Mg^{2+}	5.8	1.2	64	4.3
Na^{+}	4.0	3.3	92	8.4
K^{+}	0.7	0.8		3.5
HCO_3^-	168.0	21.0	362	72.0
SO_4^{2-}	6.4	1.2	106	6.9
Cl^-	4.8	2.4	140	3.8
Total	246.6	47.0	858	146.9

Source: R.M. Garrels and F.T. Mackenzie, *Evolution of Sedimentary Rocks* (1971). W.W. Norton and Co.

THE HYDROLOGIC CYCLE

General nature of the cycle. The present-day hydrologic cycle at the Earth's surface is illustrated in Figure 5. Some 496 cubic kilometres of water evaporate from the land and ocean surface annually, remaining for about 10 days in the atmosphere before falling as rain or snow. The amount of solar radiation necessary to evaporate this water is half of the total solar radiation received at the Earth's surface. About one-third of the precipitation falling on land runs off to the oceans primarily in rivers, while direct groundwater discharge to the oceans accounts for only about 6 percent of the total discharge. A small amount of precipitation is temporarily stored in the waters of rivers and lakes. The remaining precipitation over land, 0.073×10^6 cubic kilometres per year, returns to the atmosphere by evaporation. Over the oceans, evaporation exceeds precipitation, and the net difference represents transport of water vapour over land, where it precipitates as rain and returns to the oceans as river runoff and direct groundwater discharge.

Water residence time

The various reservoirs in the hydrologic cycle have different water residence times. Residence time is defined as the amount of water in a reservoir divided by either the rate of addition of water to the reservoir or the rate of loss from it. The oceans have a water residence time of 37,000 years; this long residence time reflects the large amount of water in the oceans. In the atmosphere the residence time of water vapour relative to total evaporation is only 10 days. Lakes, rivers, ice, and groundwaters have residence times lying between these two extremes and are highly variable.

After E.K. Berner and R.A. Berner, *The Global Water Cycle: Geochemistry and Environment,* p. 14. © 1987; reprinted by permission of Prentice Hall, Inc.

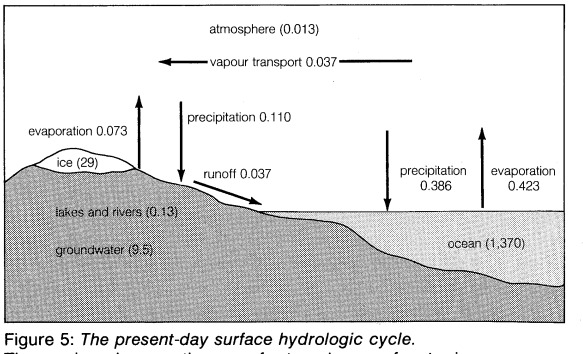


Figure 5: *The present-day surface hydrologic cycle.* The numbers in parentheses refer to volumes of water in millions of cubic kilometres, and the fluxes adjacent to the arrows are in millions of cubic kilometres of water per year.

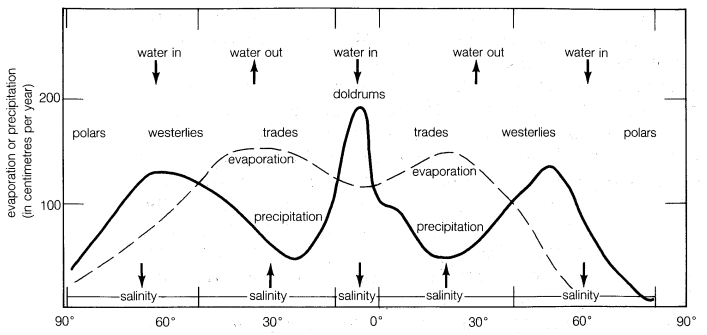


Figure 6: Latitudinal variation in precipitation and evaporation and relationship to major wind belts and oceanic salinity.

After R.M. Garrels, F.T. Mackenzie, and C. Hunt, *Chemical Cycles and the Global Environment* (1975); W. Kaufmann, Inc.

There is considerable variation in evaporation and precipitation over the globe. In order to have precipitation, there must be sufficient atmospheric water vapour and enough rising air to carry the vapour to an altitude where it can condense and precipitate. Figure 6 shows the latitudinal variation of precipitation and evaporation and their gross relation to the global wind belts. The trade winds, for example, are initially cool, but they warm up as they blow toward the equator. These winds pick up moisture from the ocean, increasing ocean surface salinity and causing seawater to sink. When the trade winds reach the equator, they rise, and the water vapour in them condenses and forms clouds. Net precipitation is high near the equator and also in the belts of the prevailing westerlies where there is frequent storm activity. Evaporation exceeds precipitation in the subtropics where the air is stable and near the poles where the air is both stable and has a low water vapour content because of the cold. The Greenland and Antarctic ice sheets formed due to the very low evaporation rates at the poles that result in precipitation exceeding evaporation in these local regions. The strong link between wind belts, the water balance, and the salinity of ocean water is apparent in Figure 6.

Processes involved in the cycle. The hydrologic cycle consists of various complicated processes of precipitation, evaporation, interception, transpiration, infiltration, percolation, retention, detention, overland flow, throughflow, and runoff. Figure 7 provides a schematic representation of the details of the hydrologic cycle, illustrating the processes involved. In the following sections, some of these processes are discussed in detail.

Water vapour and precipitation. As noted above, water exists in the atmosphere in gaseous form. Its liquid form, either as water droplets in clouds or as rain, and its solid form, as ice crystals in clouds, snowflakes, or hail, occur only momentarily and locally.

Primary functions of water vapour

Water vapour performs two major functions: (1) it is important to the radiation balance of the Earth as its presence keeps the planetary surface warmer than would otherwise be the case; and (2) it is the principal phase of the ascending part of the hydrologic cycle.

The mass of water vapour in the atmosphere, which represents only 0.001 percent of the hydrosphere, is highest in the tropics and decreases toward the poles. At a mean temperature of the Earth's surface of 15° C, the partial pressure of water vapour at equilibrium with pure water is 0.017 atmosphere. The addition of salts to pure water lowers its vapour pressure. The equilibrium, or saturation, water vapour pressure of a saturated solution of sodium chloride is 22 percent lower than that of pure water. Precipitable water vapour has, on the average, a vapour pressure of 0.0025 atmosphere, which amounts to 15 percent of the saturation vapour pressure. The ratio of observed water vapour pressure to the saturation vapour pressure at the same temperature is the relative humidity of the air. Thus, the mean relative humidity of the atmosphere is only 15 percent, a value that is low in human and biological terms. The relative humidity of the air, however, varies greatly from one geographic region to another and also vertically in the atmosphere. Atmospheric water vapour decreases rapidly with increasing altitude relative to its surface value. The amount of water required to saturate a volume of air depends on the temperature of the air. Air at high temperature can hold more water vapour at saturation than can air at low temperature. Because the temperature of the lower atmosphere (the troposphere) decreases rapidly with increasing altitude to about 15 kilometres, the upper levels of the troposphere contain little water vapour; most of the vapour is found within a few of kilometres of the Earth's surface. The average relative humidity of tropospheric air is about 50 percent. Above 15 kilometres, water vapour is essentially frozen out of

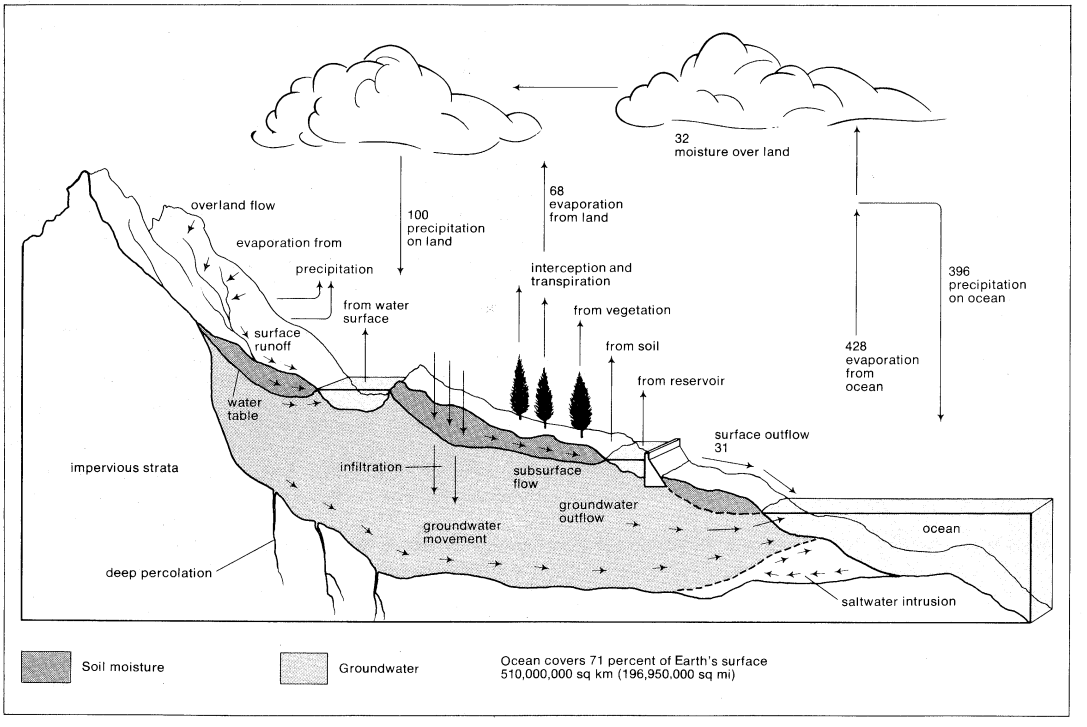


Figure 7: The detailed hydrologic cycle emphasizing processes involved in the transfer of water in the cycle. Numbers on arrows show relative water fluxes.

the atmosphere, amounting to less than 0.1 percent of its concentration at the Earth's surface.

Aside from temperature, other factors determine the water vapour content of the air and are particularly important in the lower troposphere. These factors include local evaporation and the horizontal atmospheric transportation of moisture, which varies with altitude, latitude, season, and topography. During a period of 10 days (*i.e.*, the residence time of water in the atmosphere), horizontal eddy turbulence may disperse vapour over distances up to 1,000 kilometres.

nteraction
etween
ir masses
nd surface
vater
odies

When a mass of air at the Earth's surface is exposed to a body of water, it gains water by evaporation or loses water by precipitation, depending on its relative humidity. If the air is undersaturated, with a relative humidity of less than 100 percent, it gains water vapour because the rate of evaporation exceeds the rate of condensation. If the air is supersaturated, with a relative humidity greater than 100 percent, the air mass loses water vapour because the rate of precipitation exceeds that of evaporation. This interaction between air masses and surface water bodies drives the atmosphere toward a state of saturation, which is not achieved for the entire atmosphere because of the variability in weather and because not all air masses are in contact with water bodies. In general, the level of atmospheric water vapour is higher in the summer, since temperatures are higher at this time of year. Also, atmospheric water vapour content is higher near the source of moisture than in distant regions. Over the oceans, the air is almost always near saturation, whereas over the deserts, where the supply of moisture is limited, the air is far below water vapour saturation values. In most cases, atmospheric water vapour content decreases inland over continents, but this decrease is modified by rainfall conditions, by the presence or absence of high mountains, large lakes, extensive forests, and swamps, and by the prevailing wind directions. Horizontal winds and air mass movements transfer water vapour from the ocean to the land. Although the processes are not completely separable, the horizontal transfer of water vapour seldom causes the vapour to undergo condensation, whereas vertical movements are most important in the condensation process.

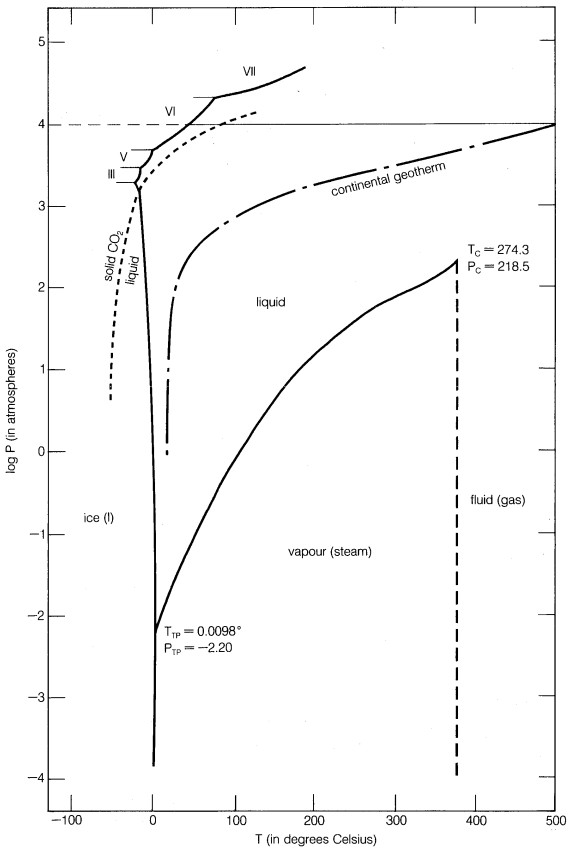


Figure 8: *The various states of water as a function of pressure and temperature.*
The continental geotherm (change of temperature with increasing depth in the Earth's crust) and the pressure–temperature boundary for solid CO_2–liquid CO_2 are shown for comparison.

Condensation depends strongly on the average temperature of the Earth's surface because the water vapour content of the air is strongly dependent on temperature. This fact is demonstrated in Figure 8, which shows the states of water as a function of the variables of pressure and temperature. The slope of the phase boundary between liquid water and water vapour is positive, implying that with increasing temperature the air at equilibrium will hold increasing amounts of water vapour. Cooling or mixing of this air results in condensation of the vapour and precipitation as water droplets or as ice crystals if the air temperature is below 0° C. When first formed, the water droplets or ice crystals are very small, on the order of 10^{-2} to 10^{-3} centimetre in diameter, and they float freely in the atmosphere. In large quantities, these water droplets and ice crystals produce a cloud. All clouds are formed as a result of cooling below the dew point, the temperature at which condensation begins when air is cooled at constant pressure and constant water vapour content. When the droplets or crystals coalesce to a size of about 10^{-2} centimetre in diameter, the drops become heavy enough to fall as raindrops or snowflakes. Hailstones measure about 10^{-1} centimetre in diameter or much larger. Water vapour condensing in the atmosphere contains strongly soluble salts (mostly of oceanic origin), weakly soluble or insoluble solids (dust), and dissolved gases. The dust and sea-salt aerosol particles in the air may act as sites of condensation by serving as nuclei for bringing initially a few water molecules together and inducing condensation from supersaturated air.

Distribution of precipitation. Precipitation falling toward the Earth's surface may suffer several fates. It may be evaporated during its fall or after it reaches the ground surface. If the surface is covered with dense vegetation, much of the precipitation may be held on leaves and plant limbs and stems. This process is termed interception and may result in little water reaching the ground because the water may be directly evaporated from plant surfaces back into the atmosphere. If precipitation reaches the ground in the form of snow, it may remain there for some time. On the other hand, if precipitation falls as rain, it may evaporate, infiltrate the soil, be detained in small catchment areas, or become overland flow—a form of runoff. Overland flow (R_o) may be expressed in terms of intensity units, water depth per unit of time (*e.g.*, centimetres per hour, or inches per hour), as

Overland flow

$$R_o = P - I, \tag{7}$$

where P is precipitation rate and I is infiltration rate (rate of entry and downward movement of water into the soil profile). Infiltration rate will equal precipitation rate until the limit of the infiltration rate, or infiltration capacity, is reached. Soil infiltration rates are usually high at the beginning of a rain preceded by a dry spell and decrease as the rainfall continues. This change in rate is due to the clogging of soil pores by particles brought from above by the infiltrating rain and to the swelling of colloidal soil particles as they absorb water. Thus, rapid decreases in infiltration rates during a rain are more likely to occur in clay-rich soils than in sandy soils.

Between rainfall periods, water held in the soil as soil moisture is gradually lost by direct evaporation or by withdrawal by plants. Evaporation into the open atmosphere occurs at the surface of the soil, and the soil dries progressively downward with time. Water vapour in the soil diffuses upward, replenishing the evaporated water, and in turn is evaporated. The pumping of air in and out of the soil by atmospheric pressure changes enhances the movement of soil moisture upward. It has been shown that evaporation of a water droplet in the free atmosphere, and to a first approximation in various soil atmospheres, is proportional to the droplet surface area $4\pi r^2$ (square centimetres), the diffusional flux of water at the droplet surface, and the transfer of heat as the droplet evaporates.

The equation for the rate of shrinkage of a water droplet due to evaporation is

$$\frac{dr}{dt} = -\frac{D\rho_{vo}(Sp-1)}{r\rho_L(1+X)}, \qquad (8)$$

where dr/dt is the rate of change in the radius of the water droplet (centimetres per second), D is the diffusion coefficient of water vapour in air (cubic centimetres per second), ρ_{vo} is the equilibrium vapour concentration at the droplet surface, Sp is the degree of undersaturation of water vapour in the environment, r is the radius of the droplet (centimetres), ρ_L is the density of liquid water (grams per cubic centimetre), and X is a dimensionless parameter depending on D, ρ_{vo}, temperature, the heat of evaporation of water vapour, the coefficient of thermal conductivity of air, and the spherical coordinate system necessary to define processes occurring to a spherical water droplet. Water droplets shrink—$dr/dt < 0$, evaporate—when the water vapour concentration in the environment (atmosphere or soil atmosphere) is less than the saturation water vapour concentration at the droplet surface. They grow—$dr/dt > 0$, condense—when the converse is true in the free atmosphere. The term dr/dt has negative values for evaporation and positive ones for condensation. Use of this equation shows, as an example, that it would take 23 minutes for a water droplet to shrink (evaporate) in size from 50 to five micrometres in air at 10° C and a water vapour undersaturation of 1 percent.

Besides simple evaporation of water from soils, water is also returned to the atmosphere by transpiration in plants. Plants draw water from soil moisture through their vast network of root hairs and rootlets. This water is carried upward through the plant trunk and branches into the leaves, where it is discharged as water vapour.

Evapotranspiration

The term evapotranspiration is used in climatic and hydrologic studies to include the combined water loss from the Earth's surface resulting from evaporation and transpiration. The maximum possible evapotranspiration is termed potential evapotranspiration and is governed by the available heat energy. It is taken as equal to evaporation from a large water surface and is generally much less than actual evapotranspiration. Actual evapotranspiration is never greater than precipitation except on irrigated land because of percolation of water into groundwater bodies and surface runoff.

Figure 9 provides a detailed picture of the hydrologic cycle in the soil zone and at depth. The soil-moisture zone gains water by precipitation and infiltration and loses water by evapotranspiration, overland flow, and percolation of water downward due to gravity into the groundwater zone. The contact between the groundwater zone (phreatic zone) and the overlying unsaturated zone (vadose zone) is called the groundwater table. The water balance equation for change of soil-moisture storage in a soil is given as

$$S = P - E - R, \qquad (9)$$

where S is storage, P is precipitation, E is evaporation, and R is surface runoff plus percolation rate into the groundwater zone; all terms are in units of length per unit of time (*e.g.*, millimetres per day, centimetres per month).

After A.N. Strahler and A.H. Strahler, *Environmental Geoscience* (1973); Hamilton Publishing Co.

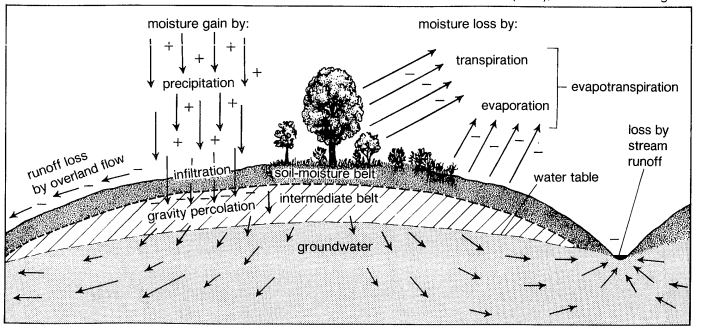


Figure 9: *The cycle of water in the soil zone and at depth.* Many factors influence water movement in this surface zone and are illustrated here.

In humid, midlatitude climates where a strong contrast between winter and summer temperatures exists, there is an annual cycle of the water content of soils. One such annual cycle is shown in Figure 10 and demonstrates the processes controlling soil moisture. Of special importance is the fact that the soils are saturated in this temperate climate in the spring, and the evaporation rate is low because of the low input of radiant energy from the Sun. By contrast, in the summer, evaporation increases because of increasing solar radiation, and with the growth of plants so does transpiration. Soil moisture is reduced to very low levels at this time of year.

After A.N. Strahler and A.H. Strahler, *Environmental Geoscience* (1973); Hamilton Publishing Co.

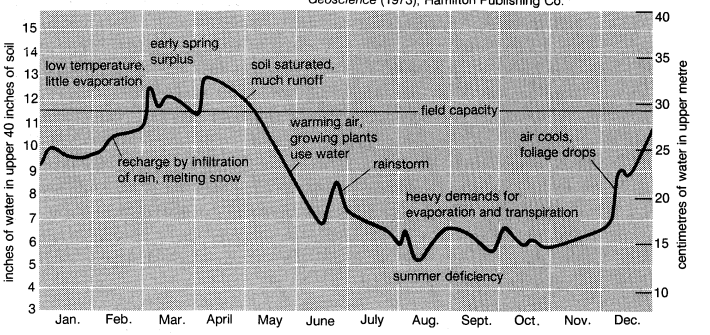


Figure 10: Annual cycle of moisture in soil in Ohio, U.S., showing factors controlling soil moisture content.

Groundwaters and river runoff. This section focuses on term R in equation (9) representing groundwater and river runoff losses from the soil-moisture zone. Water percolates from the soil-moisture zone through the unsaturated (vadose) zone to the water table. Flow through the unsaturated zone is complicated. After a rainfall, water may form nearly a continuous phase in pores in this zone, but with drying the last amount of water is held in wedges at points of contact of solid grains and as thin films on solid surfaces. The flow paths of water become more tortuous, and the water-conducting properties decrease rapidly. Structured soils and fractured rock in the vadose zone may act as conduits for fluids to reach the water table. Because of the complex geometry of water contained in the unsaturated zone, the properties of water are expressed by means of empirical relationships. Darcy's law, derived in 1856 from experimentation by the French engineer Henri Darcy, permits quantification of water flow through porous media. The law states that the rate of flow Q of a fluid through a porous layer of medium (*e.g.*, a sand bed) is directly proportional to the area A of the layer and to the difference Δh between the fluid heads at the inlet and outlet faces of the layer and inversely proportional to the thickness L of the layer, or, expressed analytically,

Darcy's law

$$Q = \frac{KA\,\Delta h}{L}, \qquad (10)$$

where K is a constant characteristic of the medium. The term K for a porous rock medium is the volume of fluid of unit viscosity passing through a unit cross section of the rock in unit time under the action of a unit pressure gradient and is called permeability. The permeability of a rock is dependent on the geometric properties of the rock, such as porosity, shape and size distribution of constituent rock grains, and degree of cementation of the rock. Permeabilities of rocks vary greatly. Unconsolidated sands may have permeabilities measured in hundreds of darcys, whereas consolidated sands that will transmit reasonable amounts of fluid have permeabilities of 0.01 to one darcy. A rough idea of the meaning of one darcy of permeability (which equals 9.869×10^{-12} square metre) can be obtained by imagining a cube of sand one foot on a side. If the sand has a permeability of one darcy, approximately one barrel of water per day will pass through the one-foot cube with a one-pound pressure head. The general equation of Darcy can be modified to express flow in both the unsaturated zone and the saturated groundwater zone.

Groundwater is constantly in motion. When a lake

or stream intersects the groundwater table, groundwater communicates directly with these bodies of water. If the groundwater table is higher than the stream or lake level, a pressure head will develop such that the groundwater flows into the water body; conversely, if the groundwater table is lower than the river or lake level, the pressure gradient induces flow into the groundwater. Most groundwater ultimately reaches the channels of surface streams and rivers and flows to the sea. On the average, groundwater contributes to total river runoff about 30 percent of its water on a global basis.

Water runoff from the land surface is that part of precipitation which eventually appears in perennial or intermittent surface streams. Streamflow-generation mechanisms have been studied for several decades, and there is now considerable knowledge regarding rainfall-runoff processes and their controls. This understanding is the result of both careful observations from field experiments and the heuristic simulations of hypothetical realities with rigorous mathematical models. The discharge measured at the downstream end of a channel reach is supplied by channel inflow at the upstream end of the reach and by the lateral inflows that enter the channel from the hillslope along the reach. The lateral inflows may arrive at the stream in one of three forms: (1) groundwater flow, (2) subsurface storm flow, or (3) overland flow.

Groundwater flow provides the base flow component of streams that sustains their flow between storms. The "flashy" response of streamflow to individual precipitation events may be ascribed to either subsurface storm flow or overland flow. Subsurface storm flow can be a dominant streamflow-generation mechanism only when the impeding subsoil horizon laterally diverts infiltrating water downslope. Under intense rainfall events during which the surface soil layer becomes saturated to some depth, water is able to migrate through "preferred pathways" rapidly enough to deliver contributions to the stream during the peak runoff period. The conditions for subsurface storm flow are quite restrictive. The mechanism is most likely to be operative on steep, humid, forested hillslopes with very permeable surface soils.

Overland flow is generated at a point on a hillslope only after surface ponding takes place. Ponding cannot occur until the surface soil layers become saturated. It is now widely recognized that surface saturation can occur because of two quite distinct mechanisms—namely, Horton overland flow and Dunne overland flow.

Horton overland flow

The former classic mechanism is for a precipitation rate that exceeds the saturated hydraulic conductivity of the surface soil. A moisture content versus depth profile during such a rainfall event will show moisture contents that increase at the surface as a function of time. At some point in time the surface becomes saturated, and an inverted zone of saturation begins to propagate downward into the soil. It is at this time that the infiltration rate drops below the rainfall rate and overland flow is generated. The time is called the ponding time. The necessary conditions for the generation of overland flow by the Horton mechanism are (1) a rainfall rate greater than the saturated hydraulic conductivity of the soil and (2) a rainfall duration longer than the required ponding time for a given initial moisture profile. Horton overland flow is generated from partial areas of the hillslope where surface hydraulic conductivities are lowest.

Dunne overland flow

In Dunne overland flow, the precipitation rate is less than the saturated hydraulic conductivity, and the initial water table is shallow or there is a shallow impeding layer. Surface saturation occurs because of a rising water table; ponding and overland flow occur at a time when no further soil-moisture storage is available. The Dunne mechanism is more common to near-channel areas. Dunne overland flow is generated from partial areas of the hillslope where water tables are shallowest. Both Horton and Dunne mechanisms result in variable source areas that expand and contract through wet and dry periods.

The distribution of total global runoff is shown in Figure 11, and the chemical characteristics of this runoff are delineated in Figure 12. It can be seen that total river discharge and the chemistry of the discharge vary from

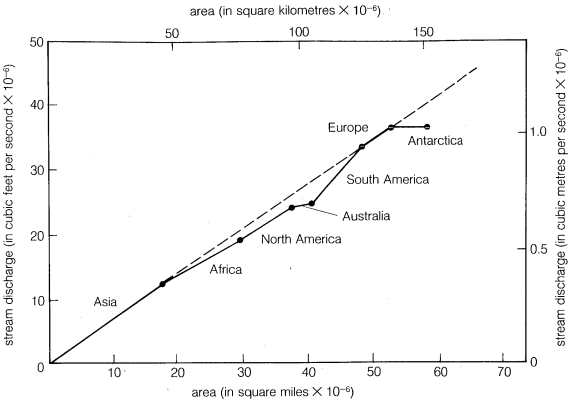


Figure 11: Distribution of total global water discharge to the oceans on a continental basis.

Adapted from *Evolution of Sedimentary Rocks* by Robert M. Garrels and Fred T. Mackenzie, by permission of W.W. Norton & Company, Inc.; copyright © 1971 by W.W. Norton & Company, Inc.

continent to continent; some continents are wetter and some drier than the world average, but the deviations are not extreme. The runoff per unit area from Asia and Europe is almost exactly equal to the world average; it is a little lower in Africa and North America; and it is considerably higher in South America. Antarctica is frozen and Australia is arid, and so they contribute little runoff. Also, since their areas are relatively small, they do not affect the global runoff average significantly. The waters draining the continents have quite different chemistries; those from Europe are very rich in calcium and bicarbonates, whereas those from Africa and South America are not. North American and Asian rivers are somewhat intermediate in their concentrations of these dissolved constituents. Such differences in composition reflect a variety of factors, including runoff, temperature, and relief, but certainly the bulk composition of the continental rocks in contact with these waters and their underground sources play a major role (see Figure 3). The surface rocks of Europe are rich in carbonates and those of South America are not; the latter are dominated by sediments rich in silicate minerals.

The chemistry of groundwater and river runoff is being modified by human activities on a global scale. The natural dissolved riverine input of major constituents to the oceans already has been increased by more than 10 percent because of human activities. In the case of sodium, chlorine, and sulfate, the increases are as high as 30 percent. In the United States alone, total water utiliza-

Adapted from *Evolution of Sedimentary Rocks* by Robert M. Garrels and Fred T. Mackenzie, by permission of W.W. Norton & Company, Inc.; copyright ©1971 by W.W. Norton & Company, Inc.

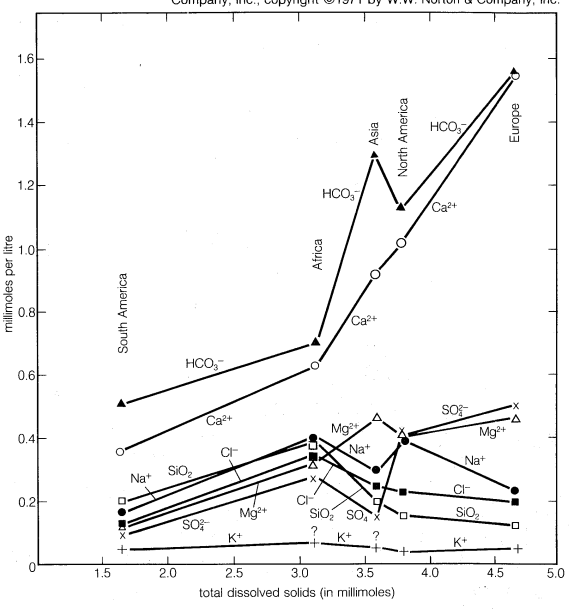


Figure 12: The chemical characteristics of river water draining various continents.

tion is equivalent to one-third of total runoff, with about 2 percent of the water used coming from underground wells. In the southwestern region of the country, water supplies have been tapped heavily and in some areas have been exhausted with no hope of replacement. This extensive utilization of fresh waters in the United States and throughout the globe make them particularly susceptible to pollution. Leachates from fertilizers, herbicides, and pesticides are found in some freshwater bodies; toxic and inorganic or organic chemicals are present; radioactive elements have been detected; and some surface-water bodies have had their salinities increased dramatically, rendering them useless for human consumption. It is therefore imperative that nations closely monitor the utilization of freshwater systems and promote their conservation.

ORIGIN AND EVOLUTION OF THE HYDROSPHERE

It is not very likely that the total amount of water at the Earth's surface has changed significantly over geologic time. Based on the ages of meteorites, the Earth is thought to be 4.6 billion years old. The oldest rocks known date 3.8 billion years in age, and these rocks, though altered by post-depositional processes, show signs of having been deposited in an environment containing water. There is no direct evidence for water for the period between 4.6 and 3.8 billion years ago. Thus, ideas concerning the early history of the hydrosphere are closely linked to theories about the origin of the Earth.

The Earth is thought to have accreted from a cloud of ionized particles around the Sun. This gaseous matter condensed into small particles that coalesced to form a protoplanet, which in turn grew by the gravitational attraction of more particulates. Some of these particles had compositions similar to that of carbonaceous chondrite meteorites, which may contain up to 20 percent water. Heating of this initially cool, unsorted conglomerate by the decay of radioactive elements and the conversion of kinetic and potential energy to heat resulted in the development of the Earth's liquid iron core and the gross internal zonation of the planet (*i.e.,* differentiation into core, mantle, and crust). It has been concluded that the Earth's core formed over a period of about 500 million years. It is likely that core formation resulted in the escape of an original primitive atmosphere and its replacement by one derived from the loss of volatile substances from the planetary interior.

Source of water for the initial hydrosphere

At an early stage the Earth thus did not have water or water vapour at its surface. Once the planet's surface had cooled sufficiently, water contained in the minerals of the accreted material and released at depth could escape to the surface and, instead of being lost to space, cooled and condensed to form the initial hydrosphere. A large, cool Earth most certainly served as a better trap for water than a small, hot body because the lower the temperature, the less likelihood for water vapour to escape, and the larger the Earth, the stronger its gravitational attraction for water vapour. Whether most of the degassing took place during core formation or shortly thereafter or whether there has been significant degassing of the Earth's interior throughout geologic time remains uncertain. It is likely that the hydrosphere attained its present volume early in the Earth's history, and since that time there have been only small losses and gains. Gains would be from continuous degassing of the Earth; the present degassing rate of juvenile water has been determined as being only 0.3 cubic kilometre per year. Water loss in the upper atmosphere is by photodissociation, the breakup of water vapour molecules into hydrogen and oxygen due to the energy of ultraviolet light. The hydrogen is lost to space and the oxygen remains behind. Only about 4.8×10^{-4} cubic kilometre of water vapour is presently destroyed each year by photodissociation. This low rate can be readily explained: the very cold temperatures of the upper atmosphere result in a cold trap at an altitude of about 15 kilometres, where most of the water vapour condenses and returns to lower altitudes, thereby escaping photodissociation. Since the early formation of the hydrosphere, the amount of water vapour in the atmosphere has been regulated by the temperature of the Earth's surface—hence its radiation balance. Higher temperatures imply higher concentrations of atmospheric water vapour, while lower temperatures suggest lower atmospheric levels.

The early hydrosphere. The gases released from the Earth during its early history, including water vapour, have been called excess volatiles because their masses cannot be accounted for simply by rock weathering. An estimate of the excess volatiles is given in Table 6. These volatiles are thought to have formed the early atmosphere of the Earth. At an initial crustal temperature of about 600° C, almost all of these compounds, including H_2O, would have been in the atmosphere. The sequence of events that occurred as the crust cooled is difficult to reconstruct. Below 100° C all of the water would have condensed, and the acid gases would have reacted with the original igneous crustal minerals to form sediments and an initial hydrosphere that was dominated by a salty ocean. If the reaction rates are assumed to have been slow relative to cooling, an atmosphere of 600° C would have contained, together with other compounds, water vapour, carbon dioxide, and hydrogen chloride (HCl) in a ratio of 20:3:1 and cooled to the critical temperature of water (*i.e.,* 374° C). The water therefore would have condensed into an early hot ocean. At this stage, the hydrogen chloride would have dissolved in the ocean (about one mole per litre), but most of the carbon dioxide would have remained in the atmosphere, with only about 0.5 mole per litre in the ocean water. This early acid ocean would have reacted vigorously with crustal minerals, dissolving out silica and cations and creating a residue composed principally of aluminous clay minerals that would form the sediments of the early ocean basins.

Formation of a hot, acid ocean

Table 6: Estimate of Excess Volatiles
(units of 10^{20} grams)

Water	16,660
Total carbon as carbon dioxide	2,500
Chlorine	300
Nitrogen	49
Sulfur	44
Boron, bromine, argon, fluorine, etc.	4

Source: J.C.G. Walker, *Evolution of the Atmosphere* (1977). Macmillan Publishing Co., Inc.

This is one of several possible pathways for the early surface of the Earth. Whatever the actual case, after the Earth's surface had cooled to 100° C, it would have taken only a short time for the remaining acid gases to be consumed in reactions involving igneous rock minerals. The presence of cyanobacteria (*e.g.,* blue-green algae) in the fossil record of rocks older than three billion years attests to the fact that the Earth's surface had cooled to temperatures lower than 100° C by this time, and neutralization of the original acid volatiles had taken place. It is possible, however, that, because of increased greenhouse gas concentrations (see below) in the Early Archean era (about 3.8 to 3.4 billion years ago), the Earth's surface could still have been warmer than today.

If most of the degassing of primary volatile substances from the Earth's interior occurred early, the chloride released by the reaction of hydrochloric acid with rock minerals would be found in the oceans or in evaporite deposits, and the oceans would have a salinity and volume comparable to that of today. This conclusion is based on the assumption that there has been no drastic change in the ratios of volatiles released through geologic time. The overall generalized reaction indicative of the chemistry leading to the formation of the early oceans can be written in the form: primary igneous rock minerals + acid volatiles + $H_2O \rightarrow$ sedimentary rocks + oceans + atmosphere. It should be noted from this equation that, if all the acid volatiles and H_2O were released early in the history of the Earth and in the proportions found today, then the total original sedimentary rock mass-produced would be equal to that of the present, and ocean salinity and volume would be close to those of today as well. If, on the other hand, degassing were linear with time, then the sedimentary rock mass would have accumulated at a linear rate, as would have oceanic vol-

ume. The salinity of the oceans, however, would remain nearly the same if the ratios of volatiles degassed did not change with time. The most likely situation is the one presented here—namely, that major degassing occurred early in Earth's history, after which minor amounts of volatiles were released episodically or continuously for the remainder of geologic time. The salt content of the oceans based on the constant proportions of volatiles released would depend primarily on the ratio of sodium chloride locked up in evaporites to that dissolved in the oceans. If all the sodium chloride in evaporites were added to the oceans today, the salinity would be approximately doubled. This value gives a sense of the maximum salinity that the oceans could have attained throughout geologic time.

)xygen roduction

One component absent from the early Earth's surface was free oxygen; it would not have been a constitutent released from the cooling crust. Early production of oxygen was by the photodissociation of water in the Earth's atmosphere, a process that was triggered by the absorption of the Sun's ultraviolet radiation. The reaction is $2H_2O \xrightarrow{h\nu} O_2 + 2H_2$, in which $h\nu$ represents the photon of ultraviolet light. The hydrogen produced would escape into space, while the oxygen would react with the early reduced gases by reactions such as $2H_2S + 3O_2 \rightarrow 2SO_2 + 2H_2O$. Oxygen production by photodissociation gave the early reduced atmosphere a start toward present-day conditions, but it was not until the appearance of photosynthetic organisms approximately three billion years ago that oxygen could accumulate in the Earth's atmosphere at a rate sufficient to give rise to today's oxygenated environment. The photosynthetic reaction leading to oxygen production is given in equation (6).

The transitional hydrosphere. The nature of the rock record from the time of the first sedimentary rocks (approximately 3.8 billion years ago) to about one to two billion years ago suggests that the amount of oxygen in the Earth's atmosphere was significantly lower than it is today and that there were continuous chemical trends in the sedimentary rocks formed and, more subtly, in the composition of the hydrosphere. Figure 13 shows how the chemistry of rocks shifted dramatically during this transitional period. The source rocks of sediments during this time may have been more basaltic than subsequent ones. Sedimentary debris was formed by the alteration of such source rocks in an oxygen-deficient atmosphere and accumulated primarily under anaerobic marine conditions. The chief difference between reactions involving mineral–ocean equilibria at this time and the present day was the role played by ferrous iron (*i.e.*, reduced state of iron). The concentration of dissolved iron in modern oceans is low because of the insolubility of oxidized iron oxides. During the transition stage and earlier, oxygen-deficient environments were prevalent, and these favoured the formation of minerals containing ferrous iron from the alteration of rocks slightly more rich in basalt than those of today. Indeed, iron carbonate siderite and iron silicate greenalite, in close association with chert and iron sulfide pyrite, are characteristic minerals that occur in iron formations of the middle Precambrian (about 2.4 to 1.5 billion years ago). The chert originally was deposited as amorphous silica; equilibrium between amorphous silica, siderite, and greenalite at 25° C and a total pressure of one atmosphere requires a carbon dioxide pressure of about $10^{-2.5}$ atmosphere, or 10 times the present-day value.

Major hange n rock hemistry

The oceans of this transitional period can be thought of as a solution that resulted from an acid leach of basaltic rocks, and, because the neutralization of the volatile acid gases was not restricted primarily to land areas as it is today, much of this alteration may have occurred by submarine processes. Anaerobic depositional environments with internal carbon dioxide pressures of about $10^{-2.5}$ atmosphere prevailed, and the oxygen-deficient atmosphere itself may have had a carbon dioxide pressure close to $10^{-2.5}$ atmosphere. If so, the pH of early ocean water was lower than that of modern seawater and the calcium concentration was higher; moreover, the early ocean water was probably saturated with respect to amorphous silica—roughly 120 ppm.

Neutral-zing of olatile cid gases

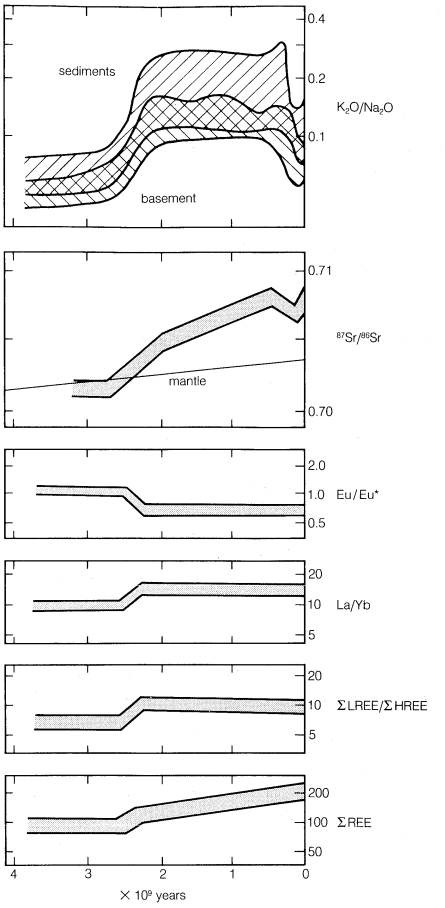


Figure 13: *Variation in the chemistry of rocks as a function of geologic age.*
There appears to be a relatively sharp transition in rock composition about two billion years ago. Eu* denotes the nonfractionated value of europium concentration; Σ LREE, Σ HREE, and Σ REE are abbreviations for total light rare earth elements, total heavy rare earth elements, and total rare earth elements, respectively.

After J. Veizer, "The Evolving Exogenic Cycle," in C.B. Gregor *et al.* (eds.), *Chemical Cycles in the Evolution of the Earth;* copyright © 1988, Wiley Interscience, New York City

To simulate what might have occurred, it is helpful to imagine emptying the Pacific basin, throwing in great masses of broken basaltic material, filling it with hydrogen chloride dissolved in water so that the acid becomes neutralized, and then carbonating the solution by bubbling carbon dioxide through it. Oxygen would not be permitted into the system. The hydrochloric acid would leach the rocks, resulting in the release and precipitation of silica and the production of a chloride ocean containing sodium, potassium, calcium, magnesium, aluminum, iron, and reduced sulfur species in the proportions present in the rocks. As complete neutralization was approached, the aluminum could begin to precipitate as hydroxides and then combine with precipitated silica to form cation-deficient aluminosilicates. As the neutralization process reached its end, the aluminosilicates would combine with more silica and with cations to form such minerals as chlorite, and ferrous iron would combine with silica and sulfur to produce greenalite and pyrite. In the final solution, chlorine would be balanced by sodium and calcium in roughly equal proportions, with subordinate amounts of potassium and magnesium; aluminum would be quantitatively removed, and silicon would be at saturation with amorphous silica. If this solution were then carbonated, calcium would be removed as calcium carbonate, and the chlorine balance would be maintained by abstraction of more sodium from the primary rock. The sediments formed in this system would contain chiefly silica, ferrous iron silicates, chloritic minerals, calcium carbonate, calcium-magnesium carbonates, and small amounts of pyrite.

If the hydrogen chloride added were in excess of the carbon dioxide, the resultant oceans would have a high content of calcium chloride ($CaCl_2$), but with a pH still near neutrality. If the carbon dioxide added were in excess of the chlorine, calcium would be precipitated as carbonate until it reached a level roughly that of present-day ocean waters—namely, a few hundred parts per million.

If this newly created ocean were left undisturbed for several hundred million years, its waters would evaporate and be transported onto the continents (in the form of precipitation); streams would transport their loads into it. The sediments produced in this ocean would be uplifted and incorporated into the continents. The influence of the continental debris would gradually be felt and the pH might change somewhat. Iron would be oxidized out of the ferrous silicates to yield iron oxides, but the composition of the water would not vary substantially.

The primary minerals of igneous rocks are all mildly basic compounds. When these minerals react in excess with acids such as hydrogen chloride and carbon dioxide, they produce neutral or mildly alkaline solutions as well as a set of altered aluminosilicate and carbonate reaction products. It is improbable that seawater has changed through time from a solution approximately in equilibrium with these reaction products—*i.e.,* with clay minerals and carbonates.

The modern hydrosphere. It is likely that the hydrosphere achieved its modern chemical characteristics about 1.5 to two billion years ago. The chemical and mineralogical compositions and the relative proportions of sedimentary rocks of this age differ little from their counterparts of the Paleozoic era (from 570 to 245 million years ago). Calcium sulfate deposits of late Precambrian age (about 1.5 billion to 570 million years ago) attest to the fact that the acid sulfur gases had been neutralized to sulfate by this time. Chemically precipitated ferric oxides in late Precambrian sedimentary rocks indicate available free oxygen, whatever its percentage. The chemistry and mineralogy of middle and late Precambrian shales are similar to those of Paleozoic shales. The carbon isotopic signature of carbonate rocks has been remarkably constant for more than three billion years, indicating exceptional stability in size and fluxes related to organic carbon. The sulfur isotopic signature of sulfur phases in rocks strongly suggests that the sulfur cycle involving heterotrophic bacterial reduction of sulfate was in operation 2.7 billion years ago. It therefore appears that continuous cycling of sediments similar to those of today has occurred for 1.5 to two billion years and that these sediments have controlled hydrospheric, and particularly oceanic, composition.

It was once thought that the saltiness of the modern oceans simply represents the storage of salts derived from rock weathering and transported to the oceans by fluvial processes. With increasing knowledge of the age of the Earth, however, it was soon realized that, at the present rate of delivery of salts to the ocean or even at much reduced rates, the total salt content and the mass of individual salts in the oceans could be attained in geologically short time intervals compared to the planet's age. The total mass of salt in the oceans can be accounted for at today's rates of stream delivery in about 12 million years. The mass of dissolved silica in ocean water can be doubled in just 20,000 years by the addition of stream-derived silica; to double the sodium content would take 70 million years. It then became apparent that the oceans were not merely an accumulator of salts; rather, as water evaporated from the oceans, together with some salt, the salts introduced must be removed in the form of minerals deposited in sediments. Accordingly, the concept of the oceans as a chemical system changed from that of a simple accumulator to that of a steady-state system in which rates of inflow of materials equal rates of outflow. The steady-state concept permits influx to vary with time, but the inflow would be matched by nearly simultaneous and equal variation of efflux.

The oceans as a steady-state system

In recent years, this steady-state conceptual view of the oceans has undergone some modification. In particular, it has been found necessary to treat components of ocean water in terms of all their influxes and effluxes and to be more cognizant of the time scale of application of the steady-state concept. Indeed, the recent increase in the carbon dioxide concentration of the atmosphere due to the burning of fossil fuels may induce a change in the pH and dissolved inorganic carbon concentrations of surface ocean water on a time scale measured in hundreds of years. If fossil-fuel burning were to cease, return to the original state of seawater composition could take thousands of years. Ocean water is not in steady state with respect to carbon on these time scales, but on a longer geologic time scale it certainly could be. Even on this longer time scale, however, oceanic composition has varied because of natural changes in the carbon dioxide level of the atmosphere and because of other factors. Figure 14 shows a model calculation of how the carbon dioxide pressure of the atmosphere and the chemistry of seawater may have varied for the past 100 million years.

From R.A. Berner, A.C. Lasaga, and R.M. Garrels in (A,D) "An Improved Geochemical Model of Atmospheric CO_2 Fluctuations Over the Past 100 Million Years," *Geophysical Monograph 32,* © 1985 by the American Geophysical Union, and (B,C) "The Carbonate–Silicate Geochemical Cycle and Its Effect on Atmospheric Carbon Dioxide Over the Past 100 Million Years," *American Journal of Science,* vol. 283 (1983)

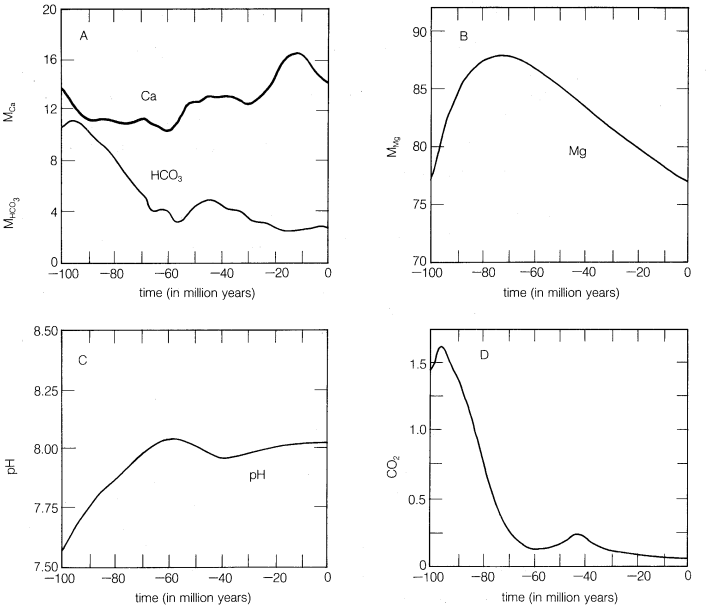


Figure 14: (A,B,C) Results of a model calculation showing possible changes in some parameters of ocean water chemistry and (D) atmospheric carbon dioxide concentration for the past 100 million years. Molar concentration is denoted by M.

It appears that the best description of modern seawater composition is that of a chemical system in a dynamic quasi-steady state. Changes in composition may occur over time, but the system always seems to return to a time-averaged, steady-state composition. In other words, since 1.5 to two billion years ago, evolutionary chemical changes in the hydrosphere have been small when viewed against the magnitude of previous change.

Table 7 shows a tentative mass balance for the modern oceans in terms of sources and sinks of some important elements. It should be noted that rivers supply dissolved constituents to the oceans, whereas high- and low-temperature reactions between seawater and submarine basalts and reactions in sediment pore waters may add or remove constituents from ocean water. Biological processes involved in the formation of the opaline silica skeletons of diatoms and radiolarians and the carbonate skeletons of planktonic foraminiferans and coccolithophorids chiefly remove calcium and silica from seawater. Exchange reactions between river-borne clays entering seawater are particularly significant for sodium and calcium ions. The mass balance of Table 7, highlighting the major processes that govern the steady-state composition of seawater on a long time scale for the elements shown, is reasonable except for sodium and iron. Most of the carbon imbalance represents carbon released to the ocean–atmosphere system during precipitation of carbonate minerals—*i.e.,*

$$Ca^{2+} + 2HCO_3^- = CaCO_3 + CO_2 + H_2O. \qquad (11)$$

Table 7: Tentative Mass Balance of Some Dissolved Species in Seawater (in units of 10^{12} grams per year)

element	river input*	hydrothermal reactions	submarine weathering	pore water	exchange reactions	biological removal	net balance
Na	+131†	−11‡	−5	−9(?)	−43		+63
Mg	+129	−60	+26	−82	−8	−6	−1
Al	+1.9				0	−2	0
Si	+203	+25	+10	+185		−375	+48
K	+50	−5	−10	−26	−8		+1
Ca	+495	+72	+32	+40	+38	−645	+32
Fe	+1.5	+15	+7	0	0	0	+23.5
C§	+389	+60	+60	+60	+5	−215	+299‖

*Corrected for cyclic salts. †The + indicates input to the ocean. ‡The − indicates output from the ocean. §The values provided here are for dissolved inorganic carbon. ‖ 215×10^{12} grams of C per year of this net imbalance for carbon represents primarily carbon released to the ocean–atmosphere system during precipitation of carbonate minerals.

Source: Adapted from R.W. Wollast and F.T. Mackenzie, "Global Cycle of Silica," in S.R. Aston (ed.), *Silicon Geochemistry and Biogeochemistry* (1983). D. Reidel Publishing Co.

In the case of iron, it has been documented that "dissolved" iron carried by rivers is rapidly precipitated as hydroxides in the mixing zone with seawater and that the reduced dissolved iron released from anaerobic sediments also is rapidly precipitated under the oxic conditions (*i.e.,* those with oxygen present) prevailing in the water column. Iron is also precipitated as iron smectites, hydrated iron oxides, and nontronite (iron-rich montmorillonite) in the deep sea. It is thus likely that iron is removed by these processes.

odium nbalance

The imbalance in sodium is large; 45 percent of the river input is not accounted for in the mass balance calculations. There are, however, major uncertainties in the estimation of the pore-water flux of sodium ions. An important sink for sodium on a geologic time scale is the formation of evaporites. If the amount of unbalanced sodium is expressed in terms of halite deposition, it would correspond to 1.6×10^{14} grams of sodium chloride per year as compared to a potential total depositional rate of 3.3×10^{14} grams annually. There are no important sodium chloride deposits forming today; thus, one possibility is that sodium is accumulating in the oceans. If so, in 6×10^{6} years at an accumulation rate of 63×10^{12} grams of sodium annually, the average salinity of the oceans would increase less than one part per thousand. The chlorine balance for the oceans, however, indicates that it is likely that the major problem in the imbalance for sodium lies in the flux estimates for sediment pore waters and perhaps submarine weathering processes.

Modern seawater chemistry has been characteristic of the past 600 million years (roughly the Phanerozoic eon) of ocean history. Evaporite sediments provide strong evidence that the composition of seawater has not varied a great deal during this interval of geologic time. Nonetheless, it seems likely, as shown in Figure 14, that fluctuations did occur, particularly in the concentrations of calcium, magnesium, and sulfate ions. The isotopic composition of sulfur in seawater, as recorded in evaporites, has varied dramatically during the past one billion years. Although it is difficult to relate these isotopic fluctuations to the calcium and sulfate concentrations of seawater, some scientists believe that the fluctuations do in fact imply changes in the latter. Furthermore, the major features of the sulfur isotopic curve for evaporites versus Phanerozoic time is similar to that of the strontium-87/strontium-86 ratio, and perhaps the strontium/calcium ratio, of sedimentary materials during this time interval. Such covariation is consistent with a model in which fluxes related to alteration of seafloor basalts and continental river runoff vary with time, resulting in variation in seawater composition.

The chemistry of the atmosphere has certainly changed significantly during the past one billion years of Earth history. A modification of this kind implies changes in the chemistry of the hydrosphere as well. Oxygen in the atmosphere rose substantially between two billion years ago and the beginning of the Phanerozoic eon (*i.e.,* 570 million years ago), whereas atmospheric carbon dioxide levels probably decreased. This change led in general to a progressively more oxygenated and less acidic hydrosphere. It is likely that the development of higher land plants during the Devonian period (from 408 to 360 million years ago) resulted in an increase in atmospheric oxygen and a decrease in carbon dioxide. Air trapped within bubbles in Arctic and Antarctic ice show that the carbon dioxide content of the atmosphere during the climax of the last ice age was about 180 parts per million by volume (ppmv), and atmospheric CO_2 levels reached approximately 280 ppmv during the last great interglacial of 120,000 years ago, long before modern society initiated its extensive fossil-fuel burning and deforestation activities (see below *Buildup of greenhouse gases*). These atmospheric changes in themselves can influence the chemistry of the hydrosphere, but they also appear to be coupled with other changes in the rock–ocean–atmosphere–biota system that strongly affect hydrospheric chemistry. For example, though surface waters probably remained oxygenated during the Cretaceous and Devonian periods of Earth history, there is evidence that intermediate and deep ocean waters were more anoxic (oxygen depleted) than today. These "anoxic events" are characterized by dramatic changes in the Earth's ocean–atmosphere system, including changes in the rates of the cyclic transfer of elements at the Earth's surface and in atmospheric composition. The probable impact of a bolide (either an asteroid or comet) and increased volcanism at the end of the Cretaceous (about 66.4 million years ago), though still the subject of hot debate, certainly could have caused short-term changes in the chemistry of the Earth's atmosphere-hydrosphere. Scientists speculate that such an event could have had any of several results, including (1) changes in the Earth's radiant-energy balance because of vast amounts of particulate and gaseous input into the atmosphere and subsequent cooling, (2) acid rain stemming from the input into the atmosphere of nitrogen gases generated by the bolide impact, (3) increased trace metal fluxes to the hydrosphere brought on by the destruction of the bolide and increased volcanism, and perhaps (4) increased "anoxia" of the hydrosphere owing to the death of land and marine organisms.

Changes in the chemistry of the Earth's atmosphere–hydrosphere

Whatever the case, evidence of change is present in the rock record albeit the composition of the modern hydrosphere has not varied greatly for the past one billion years. Moreover, as will be seen in the following section, humankind is modifying not only the chemistry of local and regional water bodies but also that of the entire global atmosphere-hydrosphere by increasing the rates of input of natural substances and by introducing new synthetic substances to the environment.

IMPACT OF HUMAN ACTIVITIES ON THE HYDROSPHERE

Introduction of contaminants

The activities of modern society are having a severe impact on the hydrologic cycle. The dynamic steady state is being disturbed by the discharge of toxic chemicals, radioactive substances, and other industrial wastes and by the seepage of mineral fertilizers, herbicides, and pesticides into surface and subsurface aquatic systems. Inadvertent and deliberate discharge of petroleum, improper sewage

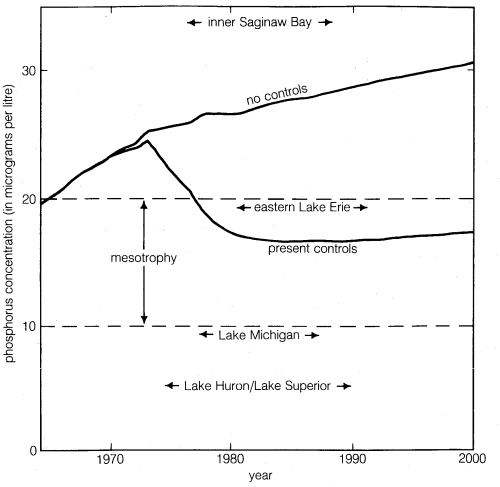


Figure 15: *Past and calculated future trends of phosphorus concentrations in the waters of Lake Ontario.* It should be noted how the present controls can bring this lake water into a state of mesotrophy (between eutrophy and oligotrophy; thus characterized by a moderate amount of dissolved nutrients). Other Great Lakes water compositions are also shown.

After S.C. Chapra, "Simulation of Recent and Projected Total Phosphorus Trends in Lake Ontario," *Journal of Great Lakes Research*, vol. 6 (1980)

disposal, and thermal pollution also are seriously affecting the quality of the hydrosphere.

The present discussion focuses on three major problems—eutrophication, acid rain, and the buildup of the so-called greenhouse gases. Each exemplifies human interference in the hydrologic cycle and its far-reaching effects.

Eutrophication. Historically, aquatic systems have been classified as oligotrophic or eutrophic. Oligotrophic waters are poorly fed by the nutrients nitrogen and phosphorus and have low concentrations of these constituents. There is thus low production of organic matter by photosynthesis (equation [6]) in such waters. By contrast, eutrophic waters are well supplied with nutrients and generally have high concentrations of nitrogen and phosphorus and, correspondingly, large concentrations of plankton owing to high biological productivity. The waters of such aquatic systems are usually murky, and lakes and coastal marine systems may be oxygen-depleted at depth. The process of eutrophication is defined as high biological productivity resulting from increased input of nutrients or organic matter into aquatic systems. For lakes, this increased biological productivity usually leads to decreased lake volume because of the accumulation of organic detritus. Natural eutrophication occurs as aquatic systems fill in with organic matter; it is distinct from cultural eutrophication, which is caused by human intervention. The latter is characteristic of aquatic systems that have been artificially enriched by excess nutrients and organic matter from sewage, agriculture, and industry. Naturally eutrophic lakes may produce 75–250 grams of carbon per square metre per year, whereas those lakes experiencing eutrophication because of human activities can support 75–750 grams per square metre per year. Commonly, culturally eutrophic aquatic systems may exhibit extremely low oxygen concentrations in bottom waters. This is particularly true of stratified systems, as, for instance, lakes during summer where concentrations of molecular oxygen may reach levels of less than about one milligram per litre—a threshold for various biological and chemical processes.

Cultural eutrophication

Aquatic systems may change from oligotrophic to eutrophic, or the rate of eutrophication of a natural eutrophic system may be accelerated by the addition of nutrients and organic matter due to human activities. The process of cultural eutrophication, however, can be reversed if the excess nutrient and organic matter supply is shut off. This response is demonstrated in Figure 15, which shows how the Great Lakes of the north central United States have responded to government regulations of the early 1970s aimed at reducing phosphorus inputs (principally from detergents) into these lakes.

Not only do freshwater aquatic systems undergo cultural eutrophication, but coastal marine systems also may be affected by this process. On a global scale, the input by rivers of organic matter to the oceans today is twice the input in prehuman times, and the flux of nitrogen, together with that of phosphorus, has more than doubled. This excess loading of carbon, nitrogen, and phosphorus is leading to cultural eutrophication of marine systems. In several polluted eastern U.S. estuaries and in some estuaries of western Europe (*e.g.*, the Scheldt of Belgium and The Netherlands), all of the dissolved silica brought into the estuarine waters by rivers is removed by phytoplankton growth (primarily diatoms) resulting from excess fluxes of nutrients and organic matter. In the North Sea, there is now a deficiency of silica and an excess of nitrogen and phosphorus, which in term has led to a decrease in diatom productivity and an increase in cyanobacteria productivity—a biotic change brought about by cultural eutrophication.

Acid rain. The emission of sulfur dioxide and nitrogen oxides to the atmosphere by human activities—primarily fossil-fuel burning—has led to the acidification of rain and freshwater aquatic systems. Acid rain is a worldwide problem and has been well documented for the eastern United States and the countries of western Europe.

Acid rain is defined as precipitation with a pH of less than 5.7 that results from reactions involving gases other than carbon dioxide. The overall reactions that produce such precipitation are those of equations (1) through (3) and

$$NO_2 + OH \rightarrow HNO_3, \qquad (12)$$

followed by dissociation of the HNO_3 to $H^+ + NO_3^-$.

Figure 16 shows the average $pH = -\log a_{H^+}$ (a_{H^+} is activity of the hydrogen ion) of precipitation over the eastern United States for the period October 1979 through September 1980. The low pH values are a result of equilibration of rainwater with the atmospheric acid gases of carbon, nitrogen, and sulfur. Equilibration only with atmospheric carbon dioxide would give a pH of 5.7. The significantly lower values are a result of reactions with nitrogen- and sulfur-bearing gaseous atmospheric components derived primarily from fossil-fuel burning sources. Nitrate and sulfate concentrations in precipitation over the eastern United States are strongly correlated with pH—the lower the pH of rain, the higher the concentrations of nitrate and sulfate. Such low pH values and increased

Reproduced, with permission, from W.D. Bischoff, V.L. Paterson, and F.T. Mackenzie, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in O.P. Bricker (ed.), *Geological Aspects of Acid Deposition* (1984); Butterworth Publishers, Stoneham, Mass.

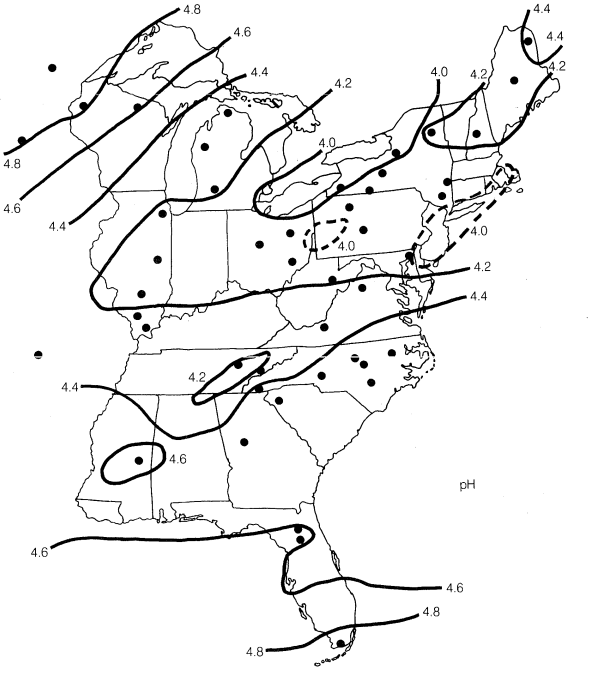


Figure 16: The average pH of precipitation during the period October 1979 to September 1980 in the eastern United States.

Table 8: Estimated Annual Fluxes of the Acid-Precipitation Constituents Sulfur and Nitrogen for the Eastern United States

flux	sulfur (10^{11} grams)	nitrogen (10^{11} grams)
A. Emission	113.9	38.8
B. Wet deposition	21.4	7.6
C. Present river	128.0	10.9
D. Pre-human river	76.8	1.4
E. Anthropogenic river (C–D)	51.2	9.5
F. Dry deposition (B–E)	29.8	1.9
G. Atmospheric transport (A–E)	62.7	29.3
H. Atmospheric transport to Canada (0.52 × grams for sulfur) (0.45 × grams for nitrogen)	32.6	13.2
I. Atmospheric transport to Atlantic (0.48 × grams for sulfur) (0.55 × grams for nitrogen)	30.1	16.1

Source: W.D. Bischoff, V.L. Paterson, and F.T. Mackenzie, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in O.P. Bricker (ed.), *Geological Aspects of Acid Deposition* (1984). Butterworth Publishers.

nitrate and sulfate concentrations also are found in the rains of western Europe and other industrialized regions of the world.

Table 8 provides an illustrative example of events occurring globally—namely, the processes—that remove anthropogenic emissions of sulfur dioxide and nitrogen oxides from the atmosphere of the eastern United States. Wet and dry deposition also removes the hydrogen ion produced in the rain by the oxidation and hydrolysis of these acid gases. This excess hydrogen ion can bring about the acidification of freshwater aquatic systems, particularly those with little buffer capacity (*e.g.*, lakes situated in crystalline rock terrains). Furthermore, the lower pH values of rainwater, and consequently of soil water, can lead to increased mobilization of aluminum. Acidification of freshwater lakes in the eastern United States and increased aluminum concentrations in their waters are thought to be responsible for major changes in the ecosystems of the lakes. In particular, many lakes of this region lack substantial fish populations today, even though they supported large numbers of fish in the early 1900s. Acid rain also may be among the factors responsible for damage to the major forests of the eastern United States and western Europe.

mpact
n lake
cosystems

Buildup of greenhouse gases. One problem brought about by human action that is definitely affecting the hydrosphere globally is that of the greenhouse gases (so called because of their heat-trapping "greenhouse" properties) emitted to the atmosphere. Of the greenhouse gases released by anthropogenic activities, carbon dioxide has received much attention. It has been shown from the measurements of carbon dioxide in air bubbles trapped in ice and from the continuous measurement of carbon dioxide concentrations in air samples collected at Mauna Loa, Hawaii, since 1958 that the present atmospheric concentration of nearly 350 ppmv is 25 percent higher than its late-1700s value. Much of this increase is due to carbon dioxide released to the atmosphere from the burning of coal, oil, gas, and wood and from the slash-and-burn activities that accompany deforestation practices (as, for example, those adopted in the Amazon River basin). The component of the hydrosphere most greatly affected by this emission of carbon dioxide is the ocean.

Figure 17A shows the manner in which carbon is cycled in the global environment on a long-term geologic basis. Before human activities had substantially affected the carbon dioxide cycle, there was a net flux of carbon dioxide from the oceans through the atmosphere to the land, where the gas was used in the net production of organic matter and the chemical weathering of minerals in continental rocks. Because of fossil-fuel burning and land-use practices, the net transfer from the ocean to the land has been reversed, and the ocean has now become an important sink of carbon dioxide (Figure 17B). The oceans are currently gaining 2,340 million tons of carbon per year. The net chemical reaction of adding carbon dioxide to the ocean (provided there is no reaction with carbonate solids) is

$$CO_2 + H_2O + CO_3^{2-} \text{ (carbonate ion)} = 2HCO_3^- \quad (13)$$

and a lowering of the pH of surface seawater. Such a pH effect has not been observed but conceivably could occur if carbon dioxide continues to be released to the atmosphere by human activities.

Based on greenhouse climate models and other considerations, it is possible that atmospheric carbon dioxide concentrations may double their late-1700s level by the years 2030–2050 and, along with those of other greenhouse gases (*e.g.*, methane and nitrous oxide), give rise to a global mean surface temperature increase of 1.5° to 5° C. This projected temperature increase would be two to three times greater at the poles than at the equator and greater in the Arctic than in the Antarctic. At present there is no worldwide program to decrease greenhouse gas emissions, except for that affecting chlorofluorocarbon (Freon) releases; thus, it is conceivable that atmospheric carbon dioxide concentrations in the late 21st and early 22nd centuries might reach levels greater than twice their 1700s value. Whatever the case, the effect of the potential rise in surface temperature would be to speed up the hydrologic cycle and probably the rate of chemical weathering of continental rocks. Increases of 4 to 7 percent in the global mean evaporation and precipitation rates might occur for a doubling of the carbon dioxide level and a few degrees rise in global mean temperature. The effect on the water balance would be regional in nature, with some places becoming wetter and others drier. In general, there would be a trend toward greater and longer periods of summer dryness induced by lower soil moisture content and higher evaporation rates in the midlatitudes of the Northern Hemisphere. In the arid western regions of the United States, which depend on irrigation for growing plants, severe water shortages could occur. By contrast, precipitation and runoff might increase, except in summer, at latitudes beyond 60° N because of a greater poleward transport of moisture. In summer, in a zone centred around 60° N, greater dryness might occur due to an earlier end of snowmelt.

Global warming and its potential effects on the hydrologic cycle

Global warming could further affect the hydrologic cycle

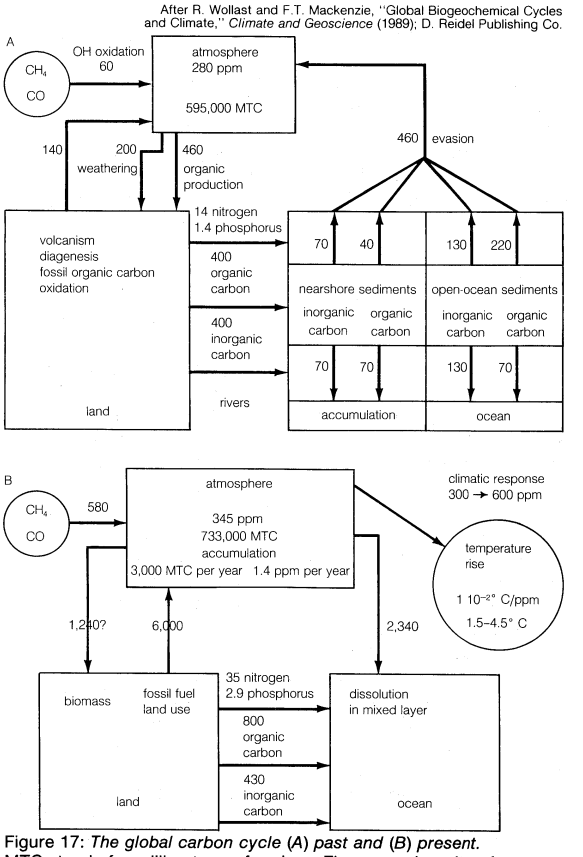


Figure 17: *The global carbon cycle (A) past and (B) present.* MTC stands for million tons of carbon. Fluxes are in units of million tons of carbon per year.

by the melting of ice and snow in the Greenland and Antarctic ice caps and in mountain glaciers, resulting in the transfer of water to the oceans. This process, together with thermal expansion of the oceans because of global warming, could lead to a slow rise in sea level of about 0.7 metre over the next century. If the West Antarctic ice sheet were to disintegrate, a much larger and more rapid rise in sea level of 5–6 metres could occur over the next several hundred years. The melting of all glacial ice would raise the sea level about 56 metres. It is also possible that a global warming could result in a reduction in the areal extent and thickness of sea ice in the Arctic and circum-Antarctic regions. Complete melting of the Arctic sea ice might occur, causing a northward shift in storm tracks and a reduction in Northern Hemispheric precipitation during the spring and fall. Furthermore, a worldwide reduction in sea ice might lead to increased evaporation from the ocean and increased low-altitude cloudiness, which would reflect solar radiation and cause cooling.

The potential changes in the hydrologic cycle induced by a global warming resulting from anthropogenic emissions of greenhouse gases do not seem great. Yet, their consequences could be severe for ecosystems and human populations, especially since the latter are so sensitive to and dependent on such changes. A global rise in sea level of one metre, for example, would almost completely inundate the coastal areas of Bangladesh. Island nations and continental beaches and cities would be endangered. Agricultural lands could be displaced, just as patterns of arid, semiarid, and wet lands might become modified. It is essential that society plan for such potential changes so that, if they do occur, appropriate adjustments can be made to accommodate them.

BIBLIOGRAPHY. General introductory discusssions on the distribution of water on and around the Earth and the role of water in supporting life are found in CYNTHIA A. HUNT and ROBERT M. GARRELS, *Water: The Web of Life* (1972); C.L. MANTELL and A.M. MANTELL, *Our Fragile Water Planet: An Introduction to the Earth Sciences* (1976); ELIZABETH KAY BERNER and ROBERT A. BERNER, *The Global Water Cycle: Geochemistry and Environment* (1987); H.M. RAGHUNATH, *Ground Water,* 2nd ed. (1987); EBERHARD CZAYA, *Rivers of the World* (1981; originally published in German, 1981); MARY J. BURGIS and PAT MORRIS, *The Natural History of Lakes* (1987); and NEIL WELLS, *The Atmosphere and Ocean: A Physical Introduction* (1986).

Biogeochemical properties of the hydrosphere are discussed in RONALD J. GIBBS, "Mechanisms Controlling World Water Chemistry," *Science* 170(3962):1088–1090 (1970); R. WOLLAST and FRED T. MACKENZIE, "Global Cycle of Silica," in S.R. ASTON (ed.), *Silicon Geochemistry and Biogeochemistry* (1983), pp.39–76; P. BUAT-MENARD, "Particle Geochemistry in the Atmosphere and Oceans," in PETER S. LISS and W. GEORGE N. SLINN (eds.), *Air-Sea Exchange of Gases and Particles* (1983), pp. 455–532; ROBERT A. BERNER, A.C. LASAGA, and ROBERT M. GARRELS, "The Carbonate-Silicate Geochemical Cycle and Its Effect on Atmospheric Carbon Dioxide Over the Past 100 Million Years," *American Journal of Science* 283(7):641–683 (1983); LAWRENCE A. HARDIE and HANS P. EUGSTER, "The Evolution of Closed-Basin Brines," *Mineralogical Society of America Special Paper* 3:273–290 (1970); JAMES I. DREVER, *The Geochemistry of Natural Waters* (1982); A. LERMAN, *Geochemical Processes: Water and Sediment Environments* (1979, reprinted 1988); G. EVELYN HUTCHINSON, *A Treatise on Limnology,* vol. 1 (1975); WERNER STUMM (ed.), *Chemical Processes in Lakes* (1985); and GEORG MATTHESS, *The Properties of Groundwater* (1982; originally published in German, 1973).

Analyses of the processes of the hydrologic cycle and its utilization are presented in ROBERT C. AVERETT and DIANE M. MCKNIGHT (eds.), *Chemical Quality of Water and the Hydrologic Cycle* (1987); R.A. FREEZE, "A Stochastic-Conceptual Analysis of Rainfall-Runoff Processes on a Hillslope," *Water Resources Research* 16(2):391–408 (1980); T. DUNNE, "Field Studies of Hillslope Flow Processes," in M.J. KIRKBY (ed.), *Hillslope Hydrology* (1978), pp. 227–293; RAY K. LINSLEY and JOSEPH B. FRANZINI, *Water-Resources Engineering,* 3rd ed. (1979); MARK J. HAMMER and KENNETH A. MACKICHAN, *Hydrology and Quality of Water Resources* (1981); and ALVIN S. GOODMAN, *Principles of Water Resources Planning* (1984).

For the evolution of the hydrosphere, see JAMES C.G. WALKER, *Evolution of the Atmosphere* (1977); J. VEIZER, "The Evolving Exogenic Cycle," in C. BRYAN GREGOR *et al.* (eds.), *Chemical Cycles in the Evolution of the Earth* (1988), pp. 175–220; HEINRICH D. HOLLAND, *The Chemical Evolution of the Atmosphere and Oceans* (1984); and ROBERT M. GARRELS and FRED T. MACKENZIE, *Evolution of Sedimentary Rocks* (1971).

The impact of human activities on the hydrosphere is studied in ROBERT M. GARRELS, FRED T. MACKENZIE, and CYNTHIA A. HUNT, *Chemical Cycles and the Global Environment: Assessing Human Influences* (1975); ARTHUR N. STRAHLER and ALAN H. STRAHLER, *Environmental Geoscience: Interaction Between Natural Systems and Man* (1973); W.D. BISCHOFF, V.L. PATERSON, and FRED T. MACKENZIE, "Geochemical Mass Balance for Sulfur- and Nitrogen-Bearing Acid Components: Eastern United States," in OWEN P. BRICKER (ed.), *Geological Aspects of Acid Deposition* (1984), pp. 1–21; S.C. CHAPRA, "Simulation of Recent and Projected Total Phosphorus Trends in Lake Ontario," *Journal of Great Lakes Research* 6(2):101–112 (1980); BRIAN HENDERSON-SELLERS and H.R. MARKLAND, *Decaying Lakes: The Origins and Control of Cultural Eutrophication* (1987); THOMAS D. BROCK, *A Eutrophic Lake: Lake Mendota, Wisconsin* (1985); G. DENNIS COOKE *et al., Lake and Reservoir Restoration* (1986); SVEN OLOF RYDING and WALTER RAST (eds.), *The Control of Eutrophication of Lakes and Reservoirs* (1989); LOUIS THIBODEAUX, *Chemodynamics, Environmental Movement of Chemicals in Air, Water, and Soil* (1979); STANLEY E. MANAHAN, *Environmental Chemistry,* 4th ed. (1984); HOWARD S. PEAVY, DONALD R. ROWE, and GEORGE TCHOBANOGLOUS, *Environmental Engineering* (1985); JAMES L. REGENS and ROBERT W. RYCROFT, *The Acid Rain Controversy* (1988); and DANIEL D. CHIRAS, *Environmental Science: A Framework for Decision Making,* 2nd ed. (1988).

(F.T.M.)

Ice and Ice Formations

Ice occurs on the Earth's continents and surface waters in a variety of forms. Most notable are the continental glaciers that cover much of Antarctica and Greenland. Smaller masses of perennial ice called ice caps occupy parts of Canada and Alaska, while mountain glaciers and piedmont glaciers are confined to more restricted areas, such as mountain valleys and the flat lands below. Other occurrences of ice on land include the different types of ground ice associated with permafrost—that is, permanently frozen soil common to very cold regions. In the oceanic waters of the polar regions, icebergs occur when large masses of ice break off from glaciers or ice shelves and drift away. The freezing of seawater in these regions results in the formation of sheets of ice known as pack ice. During the winter months similar ice bodies form on lakes and rivers in many parts of the world. The origins, characteristics, and distribution of all such ice formations are treated in this article, as are the structure and properties of ice in general. For a detailed account of the widespread occurrences of glacial ice during the Earth's past, see the articles GEOCHRONOLOGY and CLIMATE AND WEATHER. See also GEOMORPHIC PROCESSES and CONTINENTAL LANDFORMS for the effects of glaciation.

This article is divided into the following sections:

Structure and properties of ice 732
Glaciers 733
  General observations 733
    Main types of glaciers
    Distribution of glaciers
    Development of glaciers
    Importance of glacier ice
  Formation and characteristics of glacier ice 734
    Transformation of snow to ice
    Physical characteristics
  Interaction of glaciers with the environment 735
    Mass balance
    Heat balance
    Glacier flow
    Climatic changes and glacier fluctuations
  The great ice sheets 736
    Antarctic Ice Sheet
    Greenland Ice Sheet
    Mass balance of the ice sheets
    Flow of the ice sheets
    Information from deep cores
  Mountain glaciers 739
    Classification of mountain glaciers
    Surface features
    Mass balance of mountain glaciers
    Flow of mountain glaciers
    Glacier surges
    Glacier hydrology
Permafrost 742
  Origin and stability of permafrost 743
  Local thickness and ice content 744
    Types of ground ice
    Ice wedges
  Surface manifestations of permafrost and seasonally frozen ground 746
    Areas underlain by permafrost
    Features related to seasonal frost
  Problems posed by permafrost 747
Icebergs and pack ice 748
  Icebergs 748
    Formation and distribution
    Size of icebergs
    Age and melting
    Wind and current effects
    Sediment transport
    Iceberg detection and destruction
  Pack ice 752
    Formation and characteristics
    Arctic pack ice
    Antarctic pack ice
    Sea ice forecasting and reconnaissance
Ice in lakes and rivers 754
  Ice in lakes 754
    Ice formation
    Ice growth
    Ice decay
    Geographic distribution
  Ice in rivers 757
    Formation and growth
    Decay and ice jams
    Ice modification
    Geographic distribution
Bibliography 759

Structure and properties of ice

Ice has an interesting crystalline structure. The crystal lattice is made up of water molecules joined to one another by comparatively weak hydrogen bonds. The oxygen atoms occur in puckered or dimpled layers, and the atoms within each layer are arranged hexagonally. The hexagonal structure can be seen in the symmetry of a snow crystal and also by focussing the Sun's radiation into the centre of a large single crystal to cause internal melting, which may produce striking melt figures called Tyndall figures. The crystal lattice of ice has an open structure in which each oxygen atom has only four close neighbours. Thus ice has low density, and the density rises upon melting.

Although the oxygen atoms are arranged in a regular crystallographic lattice, the hydrogen atoms are arranged randomly, their positions governed only by the fact that there must be two hydrogen atoms connected by covalent bonds to each oxygen atom, and there must not be more than one hydrogen atom between any pair of neighbouring oxygen atoms (one covalent and one hydrogen bond). Because of its laminar structure, the ice crystal can deform by gliding, like a deck of cards. When this deformation occurs, however, the bonds between layers all break, and the hydrogen atoms involved in those bonds must become attached to different oxygen atoms. In doing so, they migrate within the lattice, more rapidly at higher temperatures. Sometimes not all the hydrogen atoms can reach positions in which the rules stated above are satisfied, so that a few oxygen atoms remain covalently bound to one or to three hydrogen atoms. Such oxygen atoms are the sites of electric charge. The speed of crystal deformation depends on these readjustments, which in turn are very sensitive to temperature. Thus the mechanical, thermal, and electrical properties of ice are interrelated.

Deformation and the Glen flow law

Mechanical properties. Like any other crystalline solid, ice subject to stress undergoes elastic deformation, returning to its original shape when the stress ceases. The resistance of ice to elastic elongation is about two-thirds that of lead and about $1/20$ that of steel. Of far greater importance to the study of glaciers is the susceptibility of ice to plastic deformation, or creep, which permanently alters the shape of the specimen. If a shear stress or force is applied to a sample of ice for a long time, the sample will first deform elastically and will then continue to deform plastically. This latter deformation, or flow, involves two processes: intracrystalline gliding, in which the layers within an ice crystal shear parallel to each other without destroying the continuity of the crystal lattice; and recrystallization, in which crystal boundaries change in size or

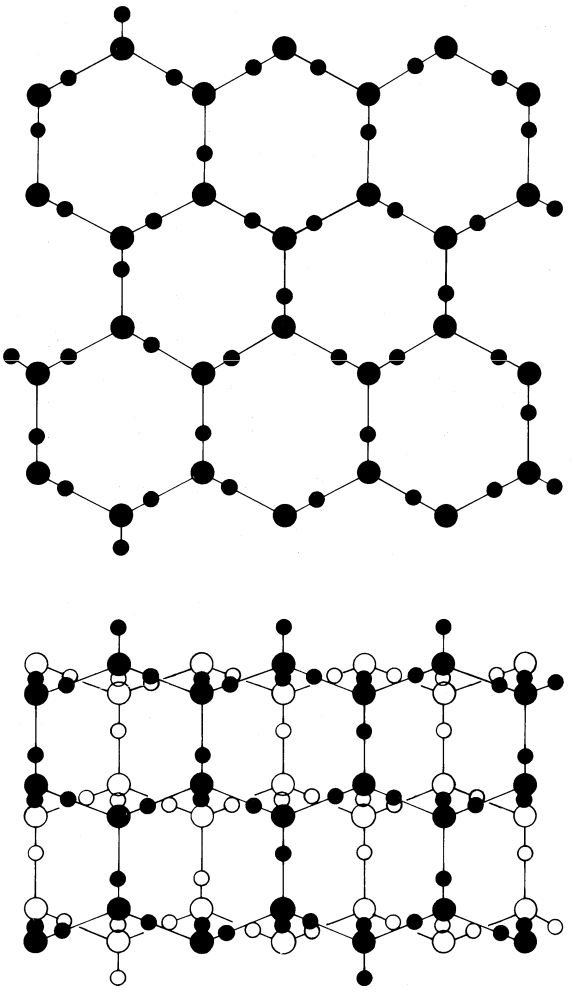


Figure 1: Atomic structure of ice. Large circles are oxygen atoms, small are hydrogen. Long lines are hydrogen bonds, short are covalent bonds. The top figure is viewed parallel to the hexagonal axis, the bottom figure perpendicular to it.
From *Proceedings of the Royal Society* (1958)

shape depending on the orientation of the adjacent crystals and the stresses exerted on them.

The rate of plastic deformation under constant shearing stress is initially high but tapers off to a steady value. If this steady value, the shear-strain rate, is plotted against the stress for many different values of applied stress, a curved graph results. The rate of shearing strain is approximately proportional to the cube of the shear stress: this is the flow law for ice. Often called the Glen flow law, it is the basis for all analyses of the flow of ice sheets and glaciers. For ice at the melting temperature, a stress of one bar causes ice to deform about 30 percent per year. If the temperature is −15° C (+5° F) the creep rate for the same stress drops to only 3 percent per year. Only shear stress is considered here; hydrostatic (equal in all directions) pressure has no effect whatever, provided the temperature is held constant relative to the freezing point.

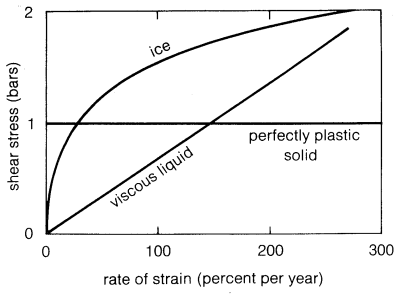


Figure 2: Flow law of ice.

This flow law implies that ice does not flow like a very viscous liquid. The strain rate of viscous liquids is directly proportional to the stress, rather than to the cube of the stress. Sometimes a "perfectly plastic" approximation is used for glacier flow in which deformation does not occur until a certain critical stress of about one bar is reached, and then the stress cannot rise above that value.

The strength of ice, which depends on many factors, is difficult to measure. If ice is stressed for a long time it deforms by plastic flow and has no yield point (at which permanent deformation begins) or ultimate strength. For short-term experiments with conventional testing machines, typical strength values in bars are 38 for crushing, 14 for bending, 9 for tensile, and 7 for shear.

Thermal properties. The heat of fusion (heat absorbed on melting of a solid) of water is 334 kilojoules per kilogram, or 79.8 calories per gram. The specific heat of ice at the freezing point is 2.04 kilojoules per kilogram per degree Celsius (0.506 calorie per degree Celsius per gram). The thermal conductivity at this temperature is 0.0226 joule per degree Celsius per second (54×10^{-4} calories per degree Celsius per second). Another property of importance to the study of glaciers is the lowering of the melting point due to hydrostatic pressure, 0.0074 degree Celsius per bar. Thus for a glacier 300 metres (984 feet) thick, everywhere at the melting temperature, the ice at the base is 0.25 degree Celsius (0.45 degree Fahrenheit) colder than at the surface.

Optical properties. Pure ice is transparent, but the air bubbles in snow and firn render it somewhat opaque. The absorption coefficient, or rate at which incident radiation decreases with depth, is about 0.1 cm^{-1} for snow and only 0.001 cm^{-1} or less for clear ice. Ice is weakly birefringent, or doubly refracting, which means that light is transmitted at different speeds in different crystallographic directions. Thin sections of snow or ice therefore can be conveniently studied under polarized light in much the same way that rocks are studied.

Albedo and radiative properties of ice

Electrical properties. The albedo, or reflectivity (an albedo of 0 means that there is no reflectivity), to solar radiation ranges from 0.5 to 0.9 for snow, 0.3 to 0.65 for firn, and 0.15 to 0.35 for glacier ice. At the thermal infrared wavelengths, snow and ice are almost perfectly "black" (absorbent), and the albedo is less than 0.01. This property means that snow and ice can either absorb or radiate long-wave radiation with high efficiency. At longer wavelengths (microwave and radio frequencies), dry snow and ice are relatively transparent. Liquid water in them greatly modifies this property. Radio echo sounding (radar) techniques are now used routinely to measure the thickness of dry polar glaciers, even where they are on the order of kilometres in thickness, but the slightest amount of liquid water distributed through the mass creates great difficulties with the technique.

Glaciers

A glacier may be defined as a large mass of perennial ice that originates on land by the recrystallization of snow or other forms of solid precipitation and that shows evidence of past or present flow. The definition is not precise, because exact limits for the terms large, perennial, and flow cannot be set. Except in terms of size, a small snow patch that persists for more than one season is hydrologically indistinguishable from a true glacier. One international group has recommended that all persisting snow and ice masses larger than 0.1 square kilometre (about 0.04 square mile) be counted as glaciers.

GENERAL OBSERVATIONS

Main types of glaciers. Glaciers are classifiable in three main groups: (1) glaciers that extend in continuous sheets, moving outward in all directions, are called ice sheets or continental glaciers if they are the size of Antarctica or Greenland and ice caps or ice fields if they are smaller; (2) glaciers confined within a path that directs the ice movement are called mountain glaciers; and (3) glaciers that spread out in cakelike sheets on level ground or on the ocean at the foot of glaciated regions are called piedmont

glaciers or shelf ice, respectively. Glaciers in the third group are not independent and are treated here in terms of their sources: ice shelves with ice sheets, piedmont glaciers with mountain glaciers.

Distribution of glaciers. A most distinctive aspect of the last ice age, the Pleistocene Epoch (from 2,500,000 to 10,000 years ago), was the recurrent expansion and contraction of the world's ice cover. These glacial fluctuations influenced geological, climatological, and biological environments and affected the evolution and development of early man. Pleistocene chronology, in fact, is based primarily on ice-sheet fluctuations. Almost all of Canada, the northern third of the United States, much of Europe, all of Scandinavia, and large parts of eastern Siberia were engulfed by ice during the major glacial stages. At times during the Pleistocene Epoch, glacial ice covered 30 percent of the world's land area; at other times the ice sheets may have shrunk to less than their present size. It may not be improper, then, to state that the world is still in an ice age. Because the term glacial generally implies ice-age events or Pleistocene time, in this discussion "glacier" is used as an adjective whenever reference is to present ice or present events.

Glacier ice today stores about three-fourths of all the fresh water in the world. Glacier ice covers about 11 percent of the world's land area, and would cause a world sea-level rise of about 90 metres (300 feet) if all existing ice melted. Glaciers occur in all parts of the world and at almost all latitudes. In Ecuador, Kenya, Uganda-Congo, and West Irian (New Guinea), glaciers even occur at the Equator, albeit at high altitudes.

ause of
acier
ıctua-
ɔns

Development of glaciers. The cause of the development and fluctuation of the world's glacier cover is still not completely understood. There is evidence that periodic fluctuations in the heat received from the Sun caused changes in ocean-surface temperature and that these also correlate with major fluctuations of glacier advance and retreat. Large ice sheets themselves, however, contain several "instability mechanisms" that may have contributed to the larger changes in world climate. One of these mechanisms, and certainly a most important one, is the very high reflectivity of dry snow to solar radiation. No other material of widespread distribution on the Earth even approaches the albedo of snow. Thus, as an ice sheet expands it causes an ever larger share of the Sun's radiation to be reflected back into space, less is absorbed on the Earth, and presumably the world's climate then cools. It has been calculated that, if the Earth's surface were covered with dry snow, the radiation balance would be changed so much that the average surface temperature would drop to −88.6° C (−128° F). Another instability mechanism is implied by the fact that the thicker and more extensive an ice sheet is, the more snowfall it will receive, due to orographic precipitation (precipitation resulting from highland effects), and the less melting will occur, due to the higher altitude of its surface and attendant lower temperature. A third possible instability mechanism has been suggested by some researchers: the Antarctic Ice Sheet may periodically surge outward and grossly extend its boundaries, thus causing a worldwide cooling and the development of ice sheets elsewhere. Although the ultimate causes of ice ages are not known with certainty, scientists agree that the world's ice cover is in a state of delicate balance, and slight external or internal changes could cause a slow but inexorable change to a new ice age or to a climate warmer and drier than now.

Importance of glacier ice. Glaciers are important to man in many other ways. In the Alps, Scandinavia, Central Asia, western Canada, and the Pacific Northwest of the United States, glacier runoff is an important source of water during dry periods and naturally regulates the flow of large rivers, thus compensating for extremes of precipitation. In the state of Washington, for instance, there is as much water temporarily stored as glacier ice as there is stored in all of the state's surface water reservoirs and lakes, and as much water is released from glaciers in July and August (the driest months) as is pumped out of the ground for a whole year.

Physical action of glacier ice

Glacier ice is an important geological agent. Mountain glacier erosion produces the spectacular U-shaped valleys, serrated ridges, and Matterhorn-like peaks that are the basis of the great beauty of Alpine scenery. The products of this erosion are deposited as moraines (ridges or sheets of glacier-derived sediment) that form a rich but rocky agricultural soil that can be dated to obtain records of past climatic events.

FORMATION AND CHARACTERISTICS OF GLACIER ICE

Transformation of snow to ice. Snow that falls on a glacier is eventually transformed to glacier ice. Many processes are involved, and the speed of transformation depends on wetness and temperature. Snow crystals are tiny hexagonal plates, needles, stars, or other intricate shapes. In a deposited snowpack these intricate shapes are usually unstable, and ice molecules evaporate off the points of crystals and are deposited in hollows or re-entrants. This causes a general rounding of the tiny ice grains so that they fit more closely together. In addition, the wind may break off the points of the intricate crystals and thus pack them more tightly. Thus, the density of the snowpack begins to increase, from its low initial value of 0.25 megagram per cubic metre (equal numerically to grams per cubic centimetre) or less, and further changes take place. Ice grains may rotate to fit more tightly together, or touching ice grains may develop necks of ice that connect them (sintering) and that grow at the expense of other parts of

The process of densification

By courtesy of the U.S. Army; photographs, A.J. Gow

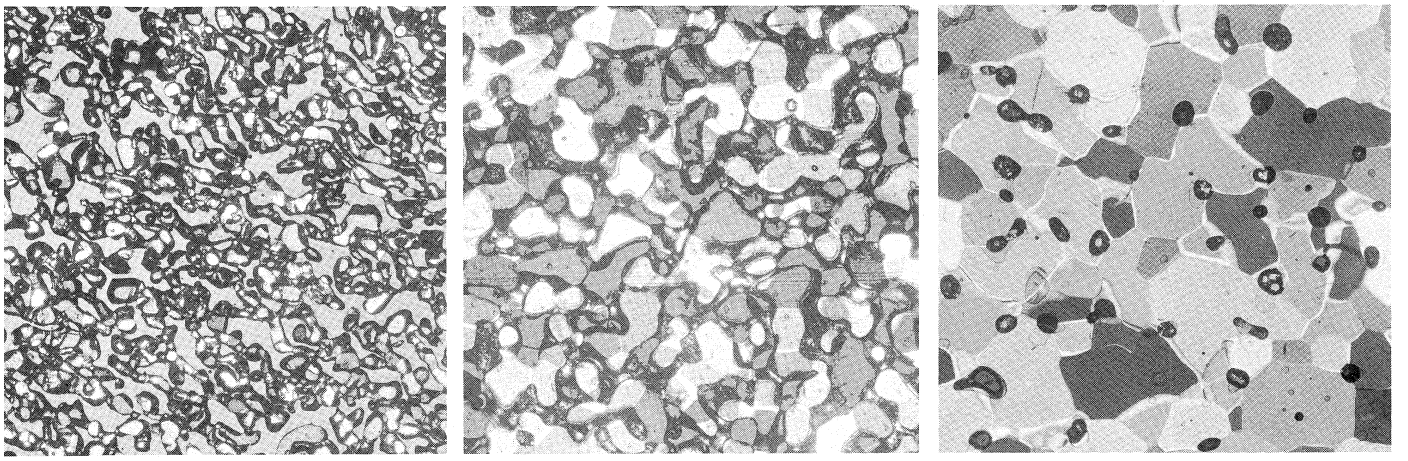


Figure 3: Thin-section photographs illustrating snow-to-ice transformation at depths of (left) 1 metre, (centre) 22 metres, and (right) 100 metres (magnified about 4.6 ×). At 1 metre and 22 metres the uniformly gray areas between grains of snow represent pore spaces. At 100 metres the snow has become ice and the bubbles become rounded and smaller because of compression of the entrapped gas. Photographs from Camp Century, Greenland, were taken between crossed polaroids to reveal crystal structure.

the ice grain. All of these processes proceed more rapidly at temperatures near the melting point and more slowly at colder temperatures.

If liquid water is present, the rate of change is many times more rapid because of the lubrication qualities of the water, the melting of ice from ice-grain extremities with refreezing elsewhere, the compacting force of surface tension, refreezing after pressure melting (regulation), and the freezing of water in air spaces.

This densification of the snow proceeds more slowly after reaching a density of 0.5–0.6 megagram per cubic metre, and many of the processes mentioned above become less and less efficient. Recrystallization becomes predominant, and grains change in size and shape in order to minimize the stress on them. This change usually means that certain favourably oriented grains grow at the expense of smaller ones. Stresses due to glacier flow may cause individual crystals to deform, and recrystallization, proceeding hand in hand with this deformation, causes an increase in the density and size of the average grain.

When the density of the aggregate reaches about 0.84 megagram per cubic metre (Mg/m^3), the air spaces between grains are sealed off, and the material becomes impermeable. Density then rises largely by the compression of air bubbles. Only rarely in mountain glaciers does the density exceed 0.90 Mg/m^3, but at great depths in ice sheets the density may approach that of pure ice (0.917 Mg/m^3 at 0° C and atmospheric pressure).

Firn (névé)

Snow that has survived one melting season is called firn (or névé); its density usually is greater than 0.5 Mg/m^3 in temperate regions but can be as low as 0.3 Mg/m^3 in polar regions. The permeability change at a density of about 0.84 Mg/m^3 marks the transition from firn to glacier ice. The transformation may take only three or four years and less than 10 metres (33 feet) of burial at South Cascade Glacier in the warm and wet environment of Washington state, but high on the plateau of Antarctica the same process takes several thousand years and burial to depths of about 150 metres (442 feet).

Physical characteristics. Glacier ice is an aggregate of irregularly shaped, interlocking single crystals that range in size from a few millimetres to several tens of centimetres. A small amount of air is present as tiny bubbles; at great depths in ice sheets the air may be dissolved in the crystals. During recrystallization, impurities migrate ahead of the grain boundary so that the interiors of large crystals are chemically very pure, especially if meltwater is involved in the process.

INTERACTION OF GLACIERS WITH THE ENVIRONMENT

Mass balance. Glaciers are nourished by snowfall, and they waste away by melting or marginal breaking (calving). In order for a glacier to remain at a constant size, there must be a balance between income (accumulation) and outgo (ablation). If this mass balance is positive (more gain than loss) the glacier grows; if it is negative the glacier shrinks.

Processes of accumulation and ablation

Accumulation refers to all processes that contribute mass to a glacier. Snowfall is the predominant process, but additional contributions may be made by hoarfrost (direct condensation of ice from water vapour), rime (freezing of supercooled water droplets on striking a surface), hail, the freezing of rain or meltwater, or avalanching of snow from adjacent slopes.

Ablation refers to all processes that remove mass from a glacier. In temperate regions, melting at the surface normally predominates. Calving (the breaking off of icebergs into the ocean or a lake) is usually the most important process on large glaciers in polar regions and on some temperate glaciers as well. Evaporation and loss by ice avalanches are important in certain special environments; floating ice may lose mass by melting from below.

Because the processes of accumulation, ablation, and the transformation of snow to ice proceed so differently, depending upon temperature and the presence or absence of liquid water, it is customary to classify glaciers in terms of various thermal conditions. Thus, a polar glacier is defined as one below the freezing temperature throughout its mass for the entire year; a subpolar glacier is mostly below the freezing temperature throughout its mass, except for surface melting in the summer; and a temperate glacier is at the melting temperature throughout its mass, but surface freezing occurs in winter.

A newer classification distinguishes the surface zones, or facies (lateral aspects), of a glacier (few single glaciers will show the full sequence). In the dry-snow zone no surface melting occurs even in summer; in the percolation zone some surface melting may occur, but the meltwater refreezes at a shallow depth so that the whole thickness of the winter snow layer is not affected; in the soaked zone sufficient melting and refreezing take place to raise the whole winter snow layer to the melting temperature; and, in the superimposed-ice zone, refrozen meltwater at the base of the snowpack (superimposed ice) forms a continuous layer that is exposed at the surface by the loss of overlying snow. These zones are all parts of the accumulation area, in which the mass balance is always positive. Below the superimposed-ice zone is the ablation zone, or ablation area, in which annual loss exceeds the gain by snowfall.

From *Journal of Glaciology*, vol. 5, no. 41 (1965); by permission of the Glaciological Society

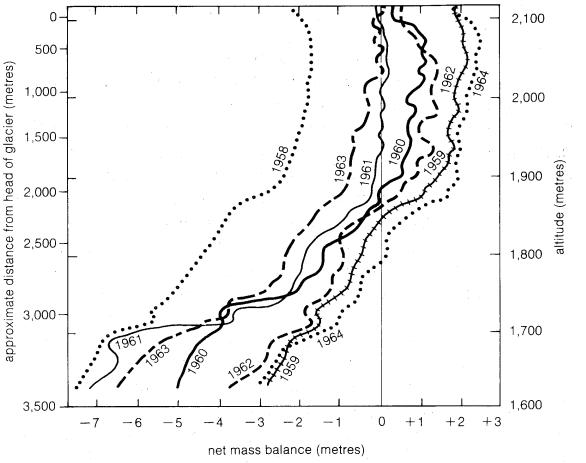


Figure 4: Net mass balance of South Cascade Glacier, Washington, 1958–64.

The boundary between the accumulation and ablation areas is the equilibrium line. Below the equilibrium line the mass balance is negative when averaged over a year. The terms firn line or firn limit are often used in place of equilibrium line on mountain glaciers, but this usage is correct only if there is no superimposed ice.

The value of the mass balance at any point on a glacier can be measured by means of stakes, snow pits, or core drills, but the change in height of the surface in relation to deeper layers and to the density of the surficial material must be observed. These values at points can then be averaged over the whole glacier for a whole year. The result is the net or annual mass balance. A positive value indicates growth, a negative value a decline.

Heat income and heat loss

Heat balance. The mass balance and the temperature variations are determined in part by the heat received from or lost to the external environment, an exchange that takes place almost entirely at the upper surface. Heat is received from solar radiation, long-wave radiation from clouds, water vapour, turbulent transfer from warm air, conduction upward from lower layers, and the heat released by the condensation of dew or hoarfrost or by the freezing of liquid water. Heat is lost by outgoing long-wave radiation, turbulent transfer to colder air, the heat required for the evaporation, sublimation, or melting of ice, and conduction downward to lower layers.

In temperate regions, solar radiation is normally the greatest heat source (although 55 to 90 percent of the incoming radiation is reflected from a snow surface), and most of the heat loss goes to the melting of ice. It is incorrect to think of snow or ice melt as directly related to air temperature; even in those situations in which heat from the air is an important factor, it is the wind structure near the surface that determines the heat transfer. This is because ordinary molecular heat conduction through the

air is extremely inefficient, and the heat can be moved to a surface only by turbulent eddies.

In polar regions, heat is gained primarily from incoming solar radiation and lost by outgoing long-wave radiation, but other processes, such as heat conduction from lower layers, turbulent transfer, and the heat of evaporation, are also involved.

Glacier flow. In the accumulation area the mass balance is positive year after year. Here the glacier would become thicker and thicker were it not for the compensating flow of ice away from the glacier. This flow also supplies mass to the ablation area to compensate for the continual loss of ice there. Glacier flow is a simple consequence of the weight and creep properties of ice. As ice tends to build up in the accumulation area, a surface slope toward the ablation area is developed. This slope and the ice's own weight induce a shearing stress throughout the mass. Each element of ice deforms according to the magnitude and direction of the shear stress and the flow law of ice. The net effect of adding up, or integrating, the sheer deformation of each increment is a velocity profile. In the simplest case, velocity decreases from the surface downward, at a rate approximately proportional to the fourth power of the depth.

Deformation and velocity distribution

When the ice temperature is entirely below freezing, it cannot slide on the underlying bedrock. If the ice is at the melting temperature in the lower layers, two mechanisms operate to permit sliding. First, small protuberances on the bed cause stress concentrations in the ice, increased plastic flow, and the ice streams around the protuberances. Second, ice on the upstream side of protuberances is subjected to higher pressure, which lowers the melting temperature and causes some of the ice to melt; on the downstream side the converse is true, and meltwater freezes. The latent heat of melting must flow upstream from the freezing side to the melting side of each bump, and the rate of heat flow partly determines the speed with which this process, termed regelation, can take place. The first process is most efficient with large knobs, and the second process is most efficient with small bedrock irregularities. Together they produce bed slip; in some situations cavities may form in the lee of bedrock knobs. Although the process of glacier sliding over bedrock is understood in a general way, none of several detailed theories has been confirmed by field observation. This problem is perhaps the most important unsolved one in glacier physics.

An area in which the velocity increases in a downglacier direction is called an area of extending flow (the separation between two stakes driven into the surface will increase with time). An area in which the velocity decreases as one moves downglacier is an area of compressing flow (the separation between stakes decreases with time). The ice itself does not actually compress or expand; the ice only deforms, and new ice is added from above or the sides, or the glacier thins in extending flow; conversely, ice is lost to the surface or goes into lateral expansion or longitudinal thickening in compressing flow.

From J.F. Nye, "A Numerical Method of Inferring Budget History of a Glacier from Its Advance and Retreat," *Journal of Glaciology*, vol. 5, no. 41 (1965); by permission of the Glaciological Society

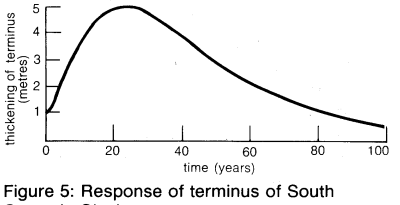


Figure 5: Response of terminus of South Cascade Glacier.

Climatic changes and glacier fluctuations. The relationship of glaciers to changes in climate is sequential. The general climatic or meteorological environment determines the local mass and heat-exchange processes at the glacier surface, and these in turn determine the net mass balance of the glacier. Changes in the net mass balance produce a dynamic response; that is, changes in the speed of ice flow. The dynamic response causes an advance or retreat of the terminus, which may produce lasting evidence of the change in the glacier margin. If the climate changes toward increased winter snowfall rates, this change makes the net mass balance more positive, which is equivalent to an increase in ice thickness. The rate of glacier flow depends on thickness in a sensitive way, so that a slight increase in thickness produces a larger increase in ice flow. This increase causes the lower margin of the glacier to push out farther before all ice is removed by melting or calving. The result is known as a glacier advance.

Glacier advance and retreat

The process, however, cannot be traced backward with assurance. A glacier advance can, perhaps, be related to a period of positive mass balances, but to ascertain the meteorological cause is difficult because either increased snowfall or decreased melting can produce a positive mass balance. As an additional complication, widespread glacier advances will add more area of high albedo to the land surface and thus tend to cause a general cooling of the atmosphere, which will, of course, promote glacier growth.

By courtesy of the U.S. Geological Survey

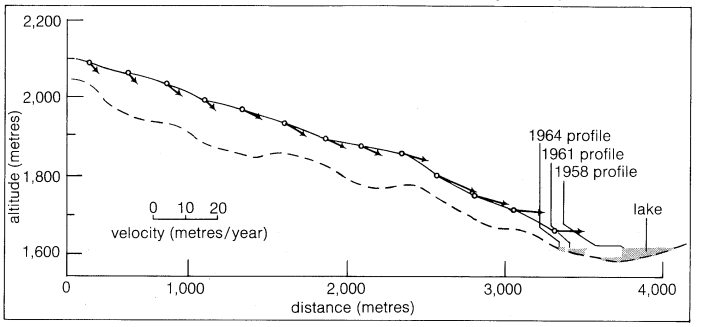


Figure 6: Profile of surface velocity vectors and calculated bedrock profiles.

The dynamic response of glaciers to changes in mass balance is better known. Although the complete equations for glacier flow are difficult to solve for changes in time, the effect of a small change or perturbation in climate can be analyzed. Such an analysis involves the theory of kinematic waves, which are akin to intermittent pulses in steady-state flow systems.

A mountain glacier or an ice sheet is first considered as a steady-state flow system, in which the addition or loss of material at any point is exactly balanced by changes in flow, so that the overall dimensions of the glacier do not change with time. Then the effect of a small or slowly varying change on the mass balance of the system is considered. A complex differential equation, which must be solved numerically, describes the effect of this perturbation on the glacier's longitudinal profile and its subsequent advance or retreat. In this way a history of mass-balance changes can be used to predict the subsequent advance or retreat of a glacier. The converse problem (determining past mass-balance changes from a known history of advance or retreat) is of more practical importance and is in actual fact far easier to do. The main problem with these analyses is the long and complex response times inherent in glaciers: even though there may be some immediate response at the terminus to change in climate, the complete response cycle may take hundreds, thousands, or even (in the case of major ice sheets) tens of thousands of years.

THE GREAT ICE SHEETS

Two great ice masses, the Antarctic and Greenland ice sheets, stand out in the world today and may be similarin many respects to the large Pleistocene ice sheets. About 99 percent of the world's glacier ice is in these two ice masses, 91 percent in Antarctica alone.

Antarctic Ice Sheet. The bedrock of the continent of Antarctica is almost completely buried under ice. Mountain ranges and isolated nunataks (a term used in Greenland and elsewhere for individual mountains surrounded by ice) locally protrude through the ice. Far greater in area are the ice shelves, where the ice sheet extends beyond the land margin and spreads out to sea. The ice sheet, with its associated ice shelves, covers an area of 13,800,000 square kilometres (5,330,000 square miles); exposed rock areas total less than 200,000 square kilometres (77,000 square

Ice topography and physical characteristics

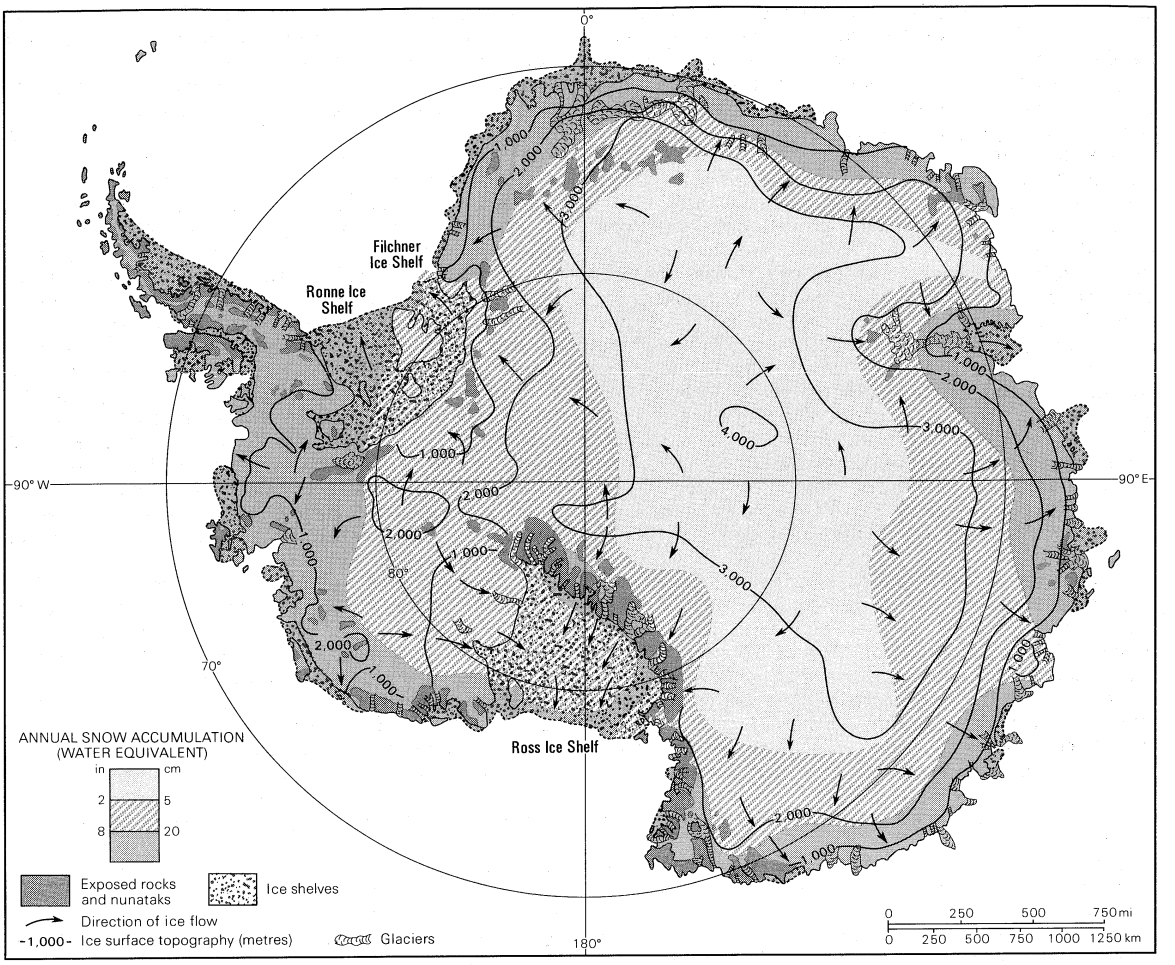


Figure 7: Nunataks, ice shelves, ice flow directions, and snow accumulation in Antarctica.
Adapted from M.B. Giovinetto, "Drainage Systems of Antarctica: Accumulation," *Antarctic Snow and Ice Studies*, Antarctic Research Series, vol. 2, American Geophysical Union (1964)

miles). The mean thickness of the ice is between 1,720 and 2,200 metres (5,600 and 7,200 feet) and the volume of ice between 25 and 30 million cubic kilometres, according to different estimates. The land surface beneath the ice is below sea level in many places, but this surface is depressed due to the weight of the ice. If the ice sheet were melted, uplift of the land surface would eventually leave only a few deep troughs and basins at or below sea level—even though the sea level itself would also rise about 80 metres (260 feet) from the addition of such a large amount of water. Because of the thick ice cover, Antarctica has by far the highest mean altitude of the continents (2.2 kilometres [1.3 miles]); all other continents have mean altitudes less than one kilometre (0.6 mile).

Antarctic Peninsula. Antarctica can be divided into three main parts: the smallest and the mildest in climate is the Antarctic Peninsula, extending from latitude 63° S off the tip of South America to a juncture with West Antarctica at a latitude of about 74° S. The ice cover of the Antarctic Peninsula is a complex of ice caps, piedmont and mountain glaciers, and small ice shelves.

West Antarctica. The part of the main continent lying south of the Americas, between 45° west longitude and 165° east longitude, is characterized by irregular bedrock and ice-surface topography and numerous nunataks and deep troughs. Two large ice shelves occur in West Antarctica: the Filchner Ice Shelf, south of the Weddell Sea, has an area of about 390,000 square kilometres (150,000 square miles), and the Ross Ice Shelf, south of the Ross Sea, has an area of about 496,000 square kilometres (191,000 square miles).

East Antarctica. A huge single ice dome of about 10,200,000 square kilometres (3,900,000 square miles), separated from West Antarctica by the Transantarctic Mountains, occurs in East Antarctica. This major mountain range extends from the eastern margin of the Ross Ice Shelf almost to the Filchner Ice Shelf. The bedrock of East Antarctica is approximately at sea level, but the ice surface locally exceeds 4,000 metres (13,000 feet) above sea level on the highest parts of the polar plateau.

Antarctic climatic conditions

At the South Pole the snow surface is 2,800 metres (9,200 feet) in altitude, and the mean annual temperature is about −50° C (−58° F), but at the Soviet Vostok Station (78° 27′ S, 106° 52′ E), 3,500 metres (11,500 feet) above sea level, the mean annual temperature is −58° C (−73° F), and in August 1960 (the winter season) the temperature reached a low of −88.3° C (−127° F). The temperatures on the polar plateau of East Antarctica are by far the coldest on Earth, and the climate of the Arctic is quite mild by comparison. Along the coast of East or West Antarctica, where the climate is milder, mean annual temperatures range from −20° to −9° C (−4° to 16° F), but temperatures exceed the melting point only for brief periods in summer, and then only slightly. Katabatic (drainage) winds, however, are very strong along the coast; the mean annual wind speed at Commonwealth Bay is 20 metres per second (45 miles per hour).

Greenland Ice Sheet. The Greenland Ice Sheet, though subcontinental in size, is huge compared with all the other glaciers in the world except that of Antarctica. Greenland (2,190,000 square kilometres [846,000 square miles]) is mostly covered by this single large ice sheet (1,730,000 square kilometres [668,000 square miles]), but isolated glaciers and small ice caps totalling 76,000 square kilometres (29,300 square miles) occur around the periphery. The ice sheet is almost 2,400 kilometres (1,500 miles) long in a north–south direction, and its greatest width is 1,100 kilometres (680 miles) at a latitude of 77° N, near its northern margin. The mean altitude of the ice surface is 2,135 metres (7,004 feet). Often the term Inland Ice, or, in Danish, Indlandsis, is used for this ice sheet.

The bedrock surface is near sea level over most of the

interior of Greenland, but mountains occur around the periphery. Thus, this ice sheet, in contrast to the Antarctic Ice Sheet, is confined along most of its margin. The ice surface reaches its greatest altitude on two north–south elongated domes, or ridges. The southern dome reaches almost 3,000 metres (9,800 feet) at latitudes 63–65° N; the northern dome reaches about 3,290 metres (10,790 feet) at about latitude 72° N. The crests of both domes are displaced east of the centre line of Greenland.

The unconfined ice sheet does not reach the sea along a broad front anywhere in Greenland, so that no large ice shelves occur. The ice margin just reaches the sea, however, in a region of many nunataks and irregular topography in the area of Melville Bay southeast of Thule. Large outlet glaciers, which are restricted tongues of the ice sheet, move through the bordering valleys to calve off into the ocean, producing the numerous icebergs that often penetrate the North Atlantic shipping lanes.

The climate of the Greenland Ice Sheet, though cold, is not as extreme as that of central Antarctica. The lowest mean annual temperatures, about −31° C (−24° F), occur on the north central part of the north dome, and temperatures at the crest of the south dome are about −20° C (−4° F).

Mass balance of the ice sheets. *Accumulation.* The rate of precipitation on the Antarctic Ice Sheet is so low that it may be called a cold desert. Snow accumulation over much of the vast polar plateau is less than five centimetres (two inches) water equivalent per year. Only around the margin of the continent, where cyclonic storms penetrate frequently, does the accumulation rise to values of more than 30 centimetres (12 inches). The mean for Antarctica is 15 centimetres (six inches) or less. In Greenland values are higher: less than 15 centimetres in a comparatively small area of north central Greenland, 30 centimetres (12 inches) along the crests of the domes, and over 80 centimetres (31 inches) along the southeast and southwest margins; the mean annual snow accumulation is about 37 centimetres (15 inches) of water equivalent.

Snow accumulation occurs mainly as direct snowfall when cyclonic storms move inland. At high altitudes on the Antarctic polar plateau, ice crystals form in the cold air during clear periods and slowly settle out as fine "diamond dust." Hoarfrost and rime deposition are generally minor items in the snow-accumulation totals. It is almost impossible to measure the precipitation directly in these climates; precipitation gauges are almost useless for the measurement of blowing snow, and the snow is blown about almost constantly in some areas. Thus, the thickness and density of snow deposited on the ground is measured, and this figure actually equals the precipitation at the place of measurement, plus the hoarfrost and rime deposition, less evaporation, less the snow blown away, and plus the snow blown in from somewhere else. The last two phenomena are thought to approximately cancel each other, except in the coastal areas, where fierce drainage, or katabatic, winds move appreciable quantities of snow.

Difficulty of measuring precipitation

The snow surface may be smooth, as in the polar areas where soft powder snow is deposited with no wind, or very hard packed and rough when high winds occur during or after snowfall. Two features are prominent: snow dunes are depositional features resembling sand dunes in their several shapes; and sastrugi are jagged erosional features (often cut into snow dunes) caused by strong prevailing winds after snowfall. Sharp, rugged sastrugi up to two metres high have been observed in high-wind areas; these make travel by vehicle or on foot very difficult. The annual snow layers exposed in the side of a snow pit can usually be distinguished by a very low density layer (depth hoar) that forms by metamorphism of the snow deposited in the fall at a time when the temperature is changing rapidly.

Most of the Antarctic Ice Sheet lies within the dry-snow zone. The percolation, soaked, and superimposed ice zones occur only in a very narrow strip in a small area along the coast. In Greenland only the central part of the northern half of the ice sheet, or about 30 percent of the total area, is within the dry-snow zone. Almost half of the area of the Greenland ice sheet is considered to be in the percolation zone. In flat areas near the equilibrium line, especially in west central Greenland, there are notorious snow swamps, or slush fields, in summer; some of this water runs off, but much of it refreezes.

Ablation. The ice sheets lose material by several processes, including surface melting, evaporation, wind erosion (deflation), calving, and the melting of the bottom surfaces of floating ice shelves by warmer seawater.

In Antarctica, calving of ice shelves and outlet glacier tongues clearly predominates among all of the processes of ice loss, but the rate of calving cannot be measured accurately at the present time. The amount of surface melt and evaporation is small, amounting to about 22 centimetres (nine inches) of ice lost from a five-kilometre (three-mile) ring around half the continent. Wind erosion is difficult to evaluate but probably accounts for only a very small loss in the mass balance. The undersides of ice shelves near their outer margins are subject to melting by the ocean water, but the rate of melting decreases inland. At some point far back under the ice shelves, freezing of seawater rather than melting must occur, but there is no agreement as to exactly where the dividing line is.

In Greenland, surface melt is more important, calving is less so, and under-shelf melting is nonexistent. Most of the calving is from the termini of a relatively few very large outlet glaciers in the middle latitudes. In Greenland and in other areas, vertical-walled melt pits in the ice are a feature of the ablation zone. Ranging from a few millimetres to a metre in diameter, these pits are floored with a dark, silty material called cryoconite, once thought to be of cosmic origin but now known to be terrestrial dust. The vertical melting of the holes is due to the absorption of solar radiation by the dark silt, possibly augmented by biological activity.

Net mass balance. Because two great ice sheets contain 99 percent of the world's ice, it might be thought that scientists would have conclusively determined whether this ice is growing or shrinking under present climatic conditions. Such is not the case. The task is simply too difficult. Because of the concentrated study of these ice sheets during and since the International Geophysical Year (1957–58), much more is known about accumulation and ablation, but some terms (such as calving) still cannot be defined satisfactorily.

Whether the ice sheets are growing or shrinking

It appears that the Antarctic Ice Sheet is growing. An important summary in 1967 listed values, in cubic kilometres of water equivalent, per year, as follows:

Accumulation	
Snow accumulation	+ 1,900
Ablation	
Melting and evaporation	− 10
Calving of ice sheet	− 50
Calving of glaciers	− 520
Calving of ice shelves	− 880
Bottom melting, ice shelves	− 200
Total ablation	− 1,660

The net difference (240 cubic kilometres), however, is of the same order of magnitude as the probable errors in estimation of the various quantities, so that a positive net balance is not a certainty. The value for accumulation is relatively lower than most, and the ablation values are relatively higher, so that the budget is a conservative one; most other modern calculations show a positive balance at least as great.

In the case of Greenland, the net balance appears to be closer to zero, but the errors are again too large for definite conclusions. The estimated balance is as follows, again in cubic kilometres of water equivalent:

accumulation		ablation	
Snow accumulation	+630	Melting	− 120 to − 270
		Calving	− 240
			− 360 to − 510

Earlier investigators had calculated an annual loss of mass, but the earlier values of melting now seem to have been grossly overestimated. Thus, the Greenland ice sheet may be growing or shrinking.

The fact that the amount of ice in the world appears to be increasing is difficult to reconcile with the observation that the sea level is also rising. Insufficient ice is lost from re-

treating mountain glaciers to make up the difference, and the rise in sea level must be due to other causes, such as a warming of the ocean or uplift of part of the ocean floor.

Flow of the ice sheets. In general, the flow in ice sheets is from central high points radially outward to the sea. In areas of subdued subglacial relief, a dome-shaped form of dynamic equilibrium is clearly apparent; from the margin the ice surface rises inland as an elliptic or parabolic curve. Wherever subglacial features are large compared with ice thickness, the ice flow is distorted and the ice surface topography is complicated. The ice of much of East Antarctica has a rather simple domal shape. Greenland resembles an elongated dome, or ridge with two summits. West Antarctica is a complex of converging and diverging flow because of the jumble of ridges and troughs in the subglacial bedrock.

Flow rates

Flow rates in the interior of an ice sheet are very low, being measured in centimetres or metres per year because the surface slope is miniscule. As the ice moves outward the rate of flow increases to a few tens to hundreds of metres per year, and this rate of flow increases still further if the flow is channeled into outlet glaciers or narrow ice shelves. Ice shelves continue the flow but in a somewhat different way—ice spreads out in ever thinner layers. This spreading, coupled with the influx of new ice from the landward side, causes the velocities to increase still more. At the edge of the Ross Ice Shelf, ice is moving out about 900 metres (3,000 feet) per year toward the ocean.

This simple picture of ice flow is made more complicated by the dependence of the flow law of ice on temperature. Because a temperature increase of about 15° C (27° F) causes a tenfold increase in the deformation rate of ice, the temperature distribution of an ice sheet partly determines its flow structure. The cold ice of the central part of an ice sheet is carried down into warmer zones. This shift modifies the static temperature distribution, and the shearing deformation is concentrated in a thin zone of warmer ice at the base. The forward velocity may be almost uniform throughout the depth to within a few tens or hundreds of metres from the bedrock.

Another important effect on ice flow is the heat produced by friction, caused by the sliding of the ice on bedrock or by internal shearing within the ice. If a portion of the ice sheet deforms more rapidly than its surroundings, the slight amount of extra heat production raises the temperature of this portion, causing it to deform even more readily. This increased deformation may explain the phenomena of ice streams—currents of ice moving several times faster than the ice on either side of them. These are very effective in moving ice from large drainage areas of Greenland out to the sea. Such ice streams have been discovered also in Antarctica, where they may be of similar importance.

Information from deep cores. Most of the Antarctic Ice Sheet is classified as polar because it is below freezing throughout its mass. It is so thick, however, that the normal gradient of temperature might be expected to increase with depth to produce melting temperatures at the base. A hole bored through 2,164 metres (7,098 feet) of ice at Byrd Station did in fact show this, and liquid water entered the hole at the base. Melting temperatures at the base probably occur in much of central Antarctica, but the thinner marginal areas are below freezing at the bed. A similar situation probably occurs in Greenland, most of which is classified as subpolar. A drill hole in northwest Greenland, where the ice is 1,390 metres (4,559 feet) thick, revealed a temperature of −13° C (+9° F) at the base.

The availability of such continuous cores through ice sheets represents one of the most exciting of modern developments. Because there is no melting, the layered structure in the ice preserves a continuous record of snow accumulation, air temperature, air composition, and artificial and natural fallout. This record extends back into the Pleistocene Epoch and is uniquely valuable for reconstructing past climates.

Near the surface it is possible to pick out the annual layers by visual inspection and quickly to detect changes in the yearly accumulations of snow. Radioactivity measurements also provide a close check on the recent layers by locating fallout from nuclear weapons. Interesting data have been obtained on the history of fallout of man-made products, such as tetraethyl lead and DDT, as well as of tiny black spherules identified as cosmic fallout.

Oxygen isotopes and paleotemperatures

A very useful technique for measuring past temperatures involves the use of oxygen isotopes, namely, oxygen-18 and oxygen-16 and their ratio. The concentration of oxygen-18 in precipitation, particularly at high latitudes, depends on the temperature. Winter snow has a more negative O^{18}–O^{16} ratio with respect to mean ocean water than summer snow. Closely spaced core samples from depths of 300 metres (1,000 feet), for example, show variations that can be clearly identified with summer and winter snows. Unfortunately, these seasonal patterns are difficult to find at great depths, because molecular diffusion tends to homogenize the isotopic variations, and the layers are progressively thinned at depth because of plastic flow.

Other techniques must be used to reconstruct a chronology from very deep cores. The most useful so far involves theoretical analysis of the flow. If the vertical profile of ice flow is known, and if it can be assumed that the rate of accumulation has been approximately constant through time, then an expression for the age of the ice as a function of depth can be developed.

This technique has been used to analyze the 1,390-metre (4,559-foot) core to bedrock from Camp Century in northwest Greenland. A plot of oxygen-18 concentration against time shows temperature variations as far back as 100,000 years, before the Wisconsin glacial period. Although these oxygen-18 changes in north Greenland cannot yet be directly converted to temperature changes, the curve agrees with known climatic variations in other parts of the world and provides far more detailed information than is available by other means. Ice cores from the dry-snow zones of ice sheets can provide continuous sedimentary records spanning perhaps several hundred millennia.

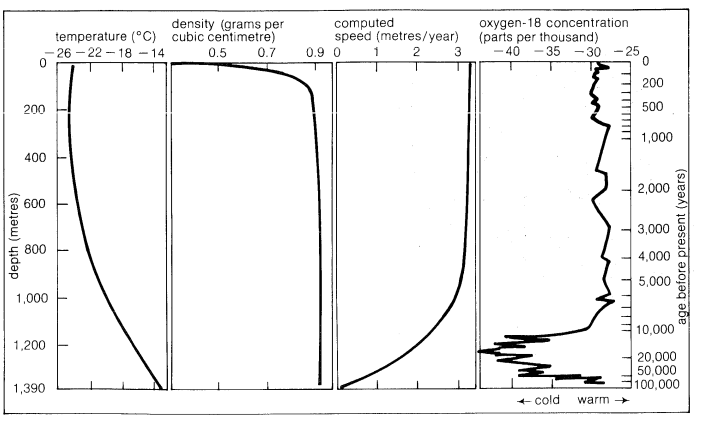


Figure 8: Data from a deep bore hole to bedrock at Camp Century, northwest Greenland.

MOUNTAIN GLACIERS

Classification of mountain glaciers. Mountain glaciers are, by definition, confined to a more or less marked path directing their movement. The shape of the channel and the degree to which the glacier fills it determine the type of glacier. Valley glaciers are of classic type; they flow at least in part down a valley and are longer than they are wide. Cirque glaciers, short and wide, are confined to cirques, or amphitheatres, cut in the mountain landscape. Other types include transection glaciers, which fill systems of valleys, and glaciers in special situations, such as summit glaciers, hanging glaciers, ice aprons, crater glaciers, and regenerated or reconstituted glaciers. Many glaciers are irregular shapes situated on mountainsides; these are often lumped into the category of wall-sided glaciers or ice patches. Included in the discussion here are glaciers that spread out at the foot of the mountain ranges (the large, composite piedmont glaciers and the smaller expanded foot glaciers) and outlet glaciers from ice sheets, ice caps, and ice fields.

Types of mountain glaciers

As the amount of glacier cover in a region increases with time, there is a gradual change from predominately cirque glaciers to valley glaciers to transection glaciers to

an ice field (almost burying the topography), with large piedmont glaciers, to an ice cap (completely covering the mountains). Thus, it is not surprising that clear distinctions among the various types of glaciers are difficult to draw. The broad spectrum of glacier types, except for the ice-cap stage, can be seen today in the St. Elias Mountains of Alaska and Yukon Territory.

Mountain glaciers are also classified as polar, subpolar, or temperate and their surfaces by the occurrence of dry-snow, percolation, saturation, and superimposed ice zones, as for ice sheets.

Surface features. The snow surface of the accumulation area of mountain glaciers displays the same snow dune and sastrugi features found on ice sheets, especially in winter, but normally these features are not as large or as well developed. Where appreciable melting of the snow occurs, several additional features may be produced. During periods of clear, sunny weather, sun cups (cup-shaped hollows usually between five and 50 centimetres [two and 20 inches] in depth) may develop. On very high-altitude, low-latitude snow and firn fields these may grow into spectacular narrow blades of ice, up to several metres high, called penitentes, or nieve penitentes. Rain falling on the snow surface (or very high rates of melt) may cause meltwater runnels (shallow grooves trending downslope).

Other features are characteristic of the ablation zone. Below icefalls (steep reaches of a valley glacier), several types of curved bands can be seen. The surface of the glacier may rise and fall in a periodic manner, with the spacing between wave crests approximately equal to the amount of ice flow in a year. Called wave ogives (pointed arches), these arcs result from the great stretching of the ice in the rapidly flowing icefall. The ice that moves through the icefall in summer has more of its surface exposed to melting and is greatly reduced in volume compared with the ice moving through in winter. The alternating thick and thin portions of the glacier appear as waves in the slower moving ice below the icefall. Dirtband ogives also occur below icefalls. These are alternating layers of dark- and light-coloured ice, which may be caused by seasonal differences in the amount of dust or by snow trapped in the icefall. Frequently these two types of ogives occur together. In plan view, the ogives are invariably distorted into arcs or curves convex downglacier; hence the name ogive.

Ogives and ice stratification

The ice of the ablation zone normally shows a distinctive layered structure. This can be relic stratification developed by the alternation of dense and light or clean and dirty snow accumulations higher on the glacier. This stratification is later subdued by recrystallization accompanying plastic flow. A new layering called foliation is developed by the flow. Foliation is expressed by alternating layers of clear and bubbly or coarse-grained and fine-grained ice. This structure has also been produced experimentally by shearing ice specimens. Although the origin of this structure is not fully understood, it is analogous to the process that produces foliated structures in metamorphic rocks (see MINERALS AND ROCKS: *Metamorphic rocks*).

The ice crystals in strongly deformed, foliated ice invariably have a preferred orientation, relative to the stress directions. In some situations, more often in polar than in temperate ice, the hexagonal axes are aligned perpendicularly to the plane of foliation. This alignment is easy to explain, because it places the crystal glide planes parallel to the planes of (presumed) greatest shearing. In most other locations the hexagonal crystal axes are preferentially aligned in four different directions, none perpendicular to the foliation. This enigmatic pattern has resisted explanation so far, in spite of much intensive experimental and theoretical study.

Crevasses are common to both the accumulation and ablation zones of mountain glaciers, as well as ice sheets and all other types of glaciers. Transverse crevasses, perpendicular to the flow direction along the centre line of valley glaciers, are caused by extending flow. Splaying crevasses, parallel to the flow in midchannel, are caused by a transverse expansion of the flow. The drag of the valley walls produces marginal crevasses, which intersect the margin at 45°. Transverse and splaying crevasses curve around to become marginal crevasses near the edge of a valley glacier. Other, less common situations produce sigmoidal (S-shaped), en echelon (overlapping), and radial crevasse patterns. Splaying and transverse crevasses may occur together, chopping the glacier surface into discrete blocks or towers, called seracs.

Crevasses deepen until the rate of surface stretching is counterbalanced by the rate of plastic flow tending to close the crevasses at depth. Thus, crevasse depths are a function of the rate of stretching and the temperature of the ice. Crevasses deeper than 50 metres (160 feet) are rare in temperate mountains, but crevasses to 100 metres (330 feet) or more in depth may occur in polar regions. Often the crevasses are concealed by a snow bridge, built by accumulations of windblown snow.

Mass balance of mountain glaciers. The rate of accumulation and ablation on mountain glaciers depends on latitude, altitude, and distance downwind from sources of abundant moisture, such as the oceans. The glaciers along the Pacific coast of Washington, British Columbia, and southeastern Alaska, as well as coastal glaciers on the South Island of New Zealand, Iceland, and southwestern Norway, receive prodigious snowfall. Snow accumulation of three to five metres (10 to 16 feet) of water equivalent

Accumulation and ablation on mountain glaciers

By courtesy of the U.S. Geological Survey; photograph, Austin Post



Figure 9: South Cascade Glacier, Washington.

in a single season is not uncommon. With this large income, glaciers can exist at low altitudes in spite of very high melt rates. The rate of snowfall increases rapidly with increasing altitude; thus, the gradient of net mass balance with altitude is steep. This gradient also expresses the rate of transfer of mass by glacier flow from high to low altitudes. The gradient of net balance at the equilibrium line is called the activity index.

Typical of the temperate, maritime glaciers is South Cascade Glacier, in western Washington. Its activity index is high, normally about 17 millimetres per metre (0.2 inch per foot; the net balance is 17 millimetres more positive for each metre gain in altitude), the yearly snow accumulation averages about 3.1 metres (10.2 feet; water-equivalent), and the equilibrium line is at the low altitude of 1,900 metres (6,200 feet). This glacier contains only ablation and saturation zones; the winter chill is so slight that no superimposed ice is formed.

In the maritime environment of southeastern Alaska are many very large glaciers; Bering and Seward-Malaspina glaciers (piedmont glaciers) cover 5,800 and 5,200 square kilometres (2,240 and 2,000 square miles) in area, respectively. Equilibrium lines are lower than those in Washington state, but the rates of accumulation and ablation and the activity indices are about the same. Because these mountains are high, and some glaciers extend over a great range of altitude, all surface zones except the dry-snow zone are represented.

Ural-type glaciers

In more continental (inland) environments, the rate of snowfall is much less, and the summer climate is generally warmer. Thus, glaciers can exist only at high altitudes. High winds may concentrate the meagre snowfall in deep, protected basins, however, allowing glaciers to form even in areas of low precipitation and high melt rates. Glaciers formed almost entirely of drift snow occur at high altitudes in Colorado and are often referred to as Ural-type glaciers because of the prevalence of this type of glacier in the polar Ural Mountains. Superimposed ice and soaked zones are found in the accumulation area. Because of the decrease in melt rates, continental glaciers in high latitudes occur at lower altitudes and have lower accumulation totals and activity indices. McCall Glacier, in the northwestern part of the Brooks Range in Arctic Alaska, has the lowest activity index (two millimetres per metre [0.02 inch per foot] measured in western North America. None of these glaciers is high enough to extend into the dry-snow zone. Glaciers in intermediate climates have intermediate equilibrium-line altitudes, accumulation or ablation totals, and activity indices.

Adapted from Meier and Tangborn, "Net Budget and Flow of South Cascade Glacier, Washington"; *Journal of Glaciology*, vol. 5, no. 41 (1965); by permission of the Glaciological Society

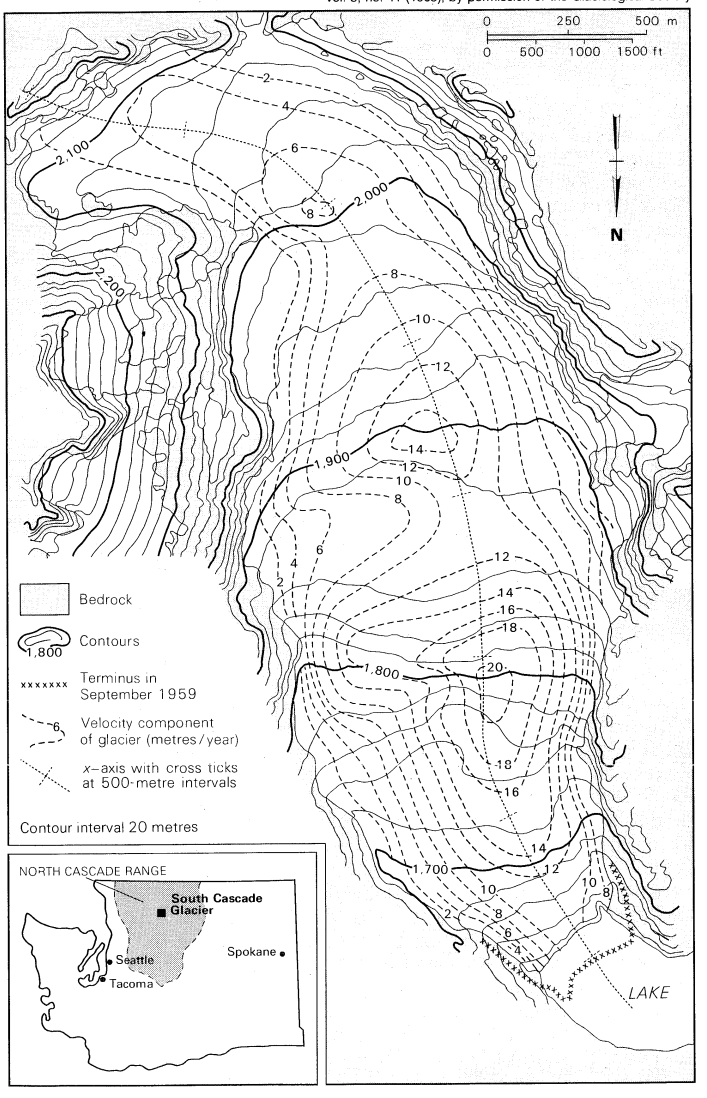


Figure 10: Topography and flow velocity of South Cascade Glacier.

Flow of mountain glaciers. Ice flow in valley glaciers has been studied extensively. The first measurements date from the mid-18th century, and the first theoretical analyses date from the middle of the 19th century. In general, these glaciers flow at rates of 0.1 to two metres (0.3 to six feet) per day, faster at the surface than at depth, faster in midchannel than along the margins, and usually faster at or just below the equilibrium line than at the head or the terminus. The flow is due, at least partly, to internal deformation, which depends primarily on the surface slope and the thickness. Glaciers that are at the melting temperature may also slide on the bed.

A common formula used to compute the speed of flow equates the speed of flow to products of ice thickness and sine of slope angle, namely,

$$u = k_1 \sin^3 \alpha\, h^4 + k_2 \sin^2 \alpha\, h^2,$$

in which u is the speed of flow at the surface, α is the surface slope, h is ice thickness, k_1 is a constant involving the shape of the channel, ice density, and other factors, and k_2 is another constant, involving a measure of the roughness of the bed. The first term on the right-hand side expresses the flow caused by internal deformation, and the second term expresses the rate of sliding on the bed. This formula is an approximation, because it ignores longitudinal changes in velocity and thickness and other complicating influences, but it has proved to be useful in simple field situations and shows that the rate of movement is very sensitive to slight changes in ice thickness.

When the thickness of a glacier is altered by changes in the net mass balance, the glacier adjusts its rate of flow to develop a different longitudinal profile. This process is normally a slow one and can be illustrated by South Cascade Glacier, Washington. It may be asked, for instance, how long it would take the glacier to overcome a minor perturbation and return to its steady state if the glacier were in complete harmony with a constant, unvarying climate, and one metre of excess thickness were added to the whole glacier for only one year. One analysis has produced the following results in terms of the behaviour of the terminus: After the initial thickening of one metre, the terminus would continue to thicken as the increased thickness propagated down from higher reaches of the glacier, and some 23 years later the glacier would be almost five metres thicker at the terminus, and then it would begin to thin. The process is so slow, however, that even after 100 years the glacier would still be 0.6 metre (two feet) thicker than at the outset. Steep, rapidly flowing glaciers should respond more rapidly and reach equilibrium more quickly, but none has been completely analyzed yet.

Glacier surges. Most glaciers follow a regular and nonspectacular pattern of advance and retreat in response to a slowly varying climate. A very different behaviour pattern has been reported for certain glaciers in several parts of the Soviet Union, Nordaustlandet (Northeastland; an island in the Arctic Ocean off the coast of Norway), Iceland, and in the Alaska Range and St. Elias Mountains of Alaska and Yukon Territory. These glaciers may, after a period of stagnation, or quiescence, lasting 10 to 100 years, suddenly begin to flow very rapidly, to up to five

metres (16 feet) per hour. This rapid flow, lasting only a year or two, causes a sudden depletion of the upper part of the glacier, accompanied by a swelling and advance of the lower part, although these usually do not reach positions beyond the limits of previous surges. Advances of several kilometres in as many months have been recorded. Even more interesting is the fact that these glaciers periodically repeat cycles of quiescence and activity, irrespective of climate. These unusual glaciers are called surging glaciers, pulsatory glaciers, catastrophically advancing glaciers, or galloping glaciers by various authorities.

Difficulty of explaining surging glaciers

Although surging glaciers are not rare in some areas (*e.g.,* Alaska Range and St. Elias Mountains), they are totally absent in other areas of similar topography, bedrock, climate, etc. (*e.g.,* western Chugach Mountains and Coast Mountains). Furthermore, glaciers of all shapes and sizes, from tiny cirque glaciers to major portions of a large ice cap, have been known to surge. Surging glaciers can be recognized readily during the active phase. Wildly shattered surfaces, sheer cliffs of ice where the main glacier has separated from tributary ice in the upper parts, and bulging masses of broken ice riding over older moraines and stagnant ice in the lower parts are characteristic. Usually they are easy to identify during the quiescent phase also, because medial moraines (central ridges of rock debris) on the surface are looped and zigzagged in a distinctive way. The flow instability that permits glacier surges has something to do with an abrupt decoupling of the glacier from its bed, but beyond that statement it is still impossible to explain glacier surges. Surging glaciers represent one of the most interesting enigmas in present-day glaciology.

Glacier hydrology. A temperate glacier is essentially a reservoir that gains precipitation in both liquid and solid form, stores a large share of this precipitation, and then releases it with little loss at a later date. The hydrologic characteristics of this reservoir, however, are complex, because its physical attributes change during a year.

In late spring the glacier is covered entirely by a thick snowpack at the melting temperature. Meltwater and liquid precipitation must travel through the snowpack by slow percolation until reaching well-defined meltwater channels in the solid ice below. In summer the snowpack becomes thinner, and drainage paths within the snow are more defined. Bare ice is exposed, and on this there may be surface drainage. Thus, meltwater and liquid precipitation are transmitted through the glacier rapidly. In winter, snow accumulates, and the surface layer freezes. This condition completely stops the movement of surface meltwater and precipitation, and rain that falls on the frozen surface can only refreeze to join the ice reservoir.

Runoff from glacier ice

The runoff from a typical Northern Hemisphere temperate glacier reaches a peak in late July or early August. Solar radiation, the chief source of heat to promote melt, reaches a peak in June. The delay in the peak melt rates is primarily because of the changing albedo (surface reflectivity) during the summer; initially the snow is very reflective and covers the whole glacier, but as the summer wears on the snow becomes wet (less reflective), and in addition more and more of the ice of much lower albedo is exposed. Thus, even though the incoming radiation decreases during midsummer, the proportion of it that is absorbed to cause melt is greatly increased. Other heat-exchange processes, such as turbulent transfer from warm air, also become more important in midsummer and late summer.

This albedo variation produces a runoff "buffering effect" against unusually wet or dry years. An unusually heavy winter snowpack causes high-albedo snow to persist longer over the glacier in summer; thus, less meltwater is produced. Conversely, an unusually light winter snowfall causes older firn and ice of lower albedo to be exposed earlier in the summer, producing increased melt and runoff. This variation can be shown by comparing data from two different years at South Cascade Glacier. During 1957–58 the spring snowpack was abnormally low (2.31 metres [eight feet] of water, 77 percent of average), and the summer runoff was therefore high (5.76 metres [18.9 feet], 158 percent of average). By way of contrast, during 1963–64, the spring snowpack was high (3.53 metres [11.6 feet], 117 percent of average), and the runoff was low (2.31 metres, 68 percent of average).

Thus, glaciers naturally regulate the runoff, seasonally and from year to year. When glacier runoff is combined with nonglacier runoff in roughly equal amounts, the result is very stable and even streamflow. This condition is part of the basis for the extensive hydrologic development of regions such as western Washington.

Glacier streams are characterized by high sediment concentrations. The sediment ranges from boulders to the distinctive fine-grained material called rock flour, glacier flour, or glacier milk, which is colloidal in size (often less than one micron in diameter). The suspended sediment concentration is a function of distance from the glacier, but the rock-flour component may persist for great distances and remain suspended in lakes for many years; it is rock flour that is responsible for the green colour of Alpine lakes. Glacier erosion has had little direct observational study, and so not much can be said about how much sediment is produced from a glacier of a given size or activity. Glacier streams vary in discharge with the time of day, and this variation causes a continual readjustment of the stream channel and the transportation of reworked debris, adding to the sediment load.

Glacier floods. Glacier outburst floods, or *jøkulhlaups,* can be spectacular or even catastrophic. These happen when drainage within a glacier is blocked by internal plastic flow and water is stored in or behind the glacier. The water eventually finds a narrow path to trickle out. If any excess heat is available in the water, this movement will cause the path to be enlarged by melting, causing faster flow, more melting, a larger conduit, and so on until all of the water is released quite suddenly. The word *jøkulhlaup* is Icelandic in origin, and Iceland has experienced some of the world's most spectacular outburst floods. The 1922 Grimsvötn outburst released about 7.1 cubic kilometres (1.7 cubic miles) of water in a flood that was estimated to have reached almost 57,000 cubic metres per second (2,000,000 cubic feet per second). Lake George (Knik Glacier), near Anchorage, and Tulsequah Glacier, near Juneau, Alaska; Summit Lake (Salmon Glacier), near Prince Rupert, British Columbia; and Kautz and Nisqually glaciers, at Mt. Rainier, Washington, are sites of outburst floods. Many outburst floods occur regularly each year, some at intervals of two or more years, and some are completely irregular and impossible to predict.

(M.F.M.)

Permafrost

Most permafrost is consolidated by ice. Some, however, contains no water, and thus no ice, and is termed dry permafrost. The upper surface of permafrost is called the permafrost table. In permafrost areas the surface layer of ground that freezes in the winter (seasonally frozen ground) and thaws in summer is called the active layer. The thickness of the active layer depends mainly on the moisture content, varying from less than a foot in thickness in wet, organic sediments to several feet in well-drained gravels.

Dry permafrost

Permafrost is widespread in the northern part of the Northern Hemisphere (Figure 11), and it is estimated to underlie 20 percent of the world's land surface. It occurs in 85 percent of Alaska, 50 percent of the U.S.S.R. and Canada, and probably all of Antarctica. In the Northern Hemisphere, permafrost is more widespread and extends to greater depths in the north than in the south. It is 1,500 metres (5,000 feet) thick in northern Siberia, 600 metres (2,000 feet) thick in northern Alaska, and thins progressively toward the south (Figure 12). Permafrost can be differentiated into two broad zones; the continuous and the discontinuous, referring to the lateral continuity of permafrost. In the continuous zone of the far north, permafrost is nearly everywhere present except under the lakes and rivers that do not freeze to the bottom. The discontinuous zone includes numerous permafrost-free areas that increase progressively in size and number from north to south. Near the southern boundary, only rare patches of permafrost exist.

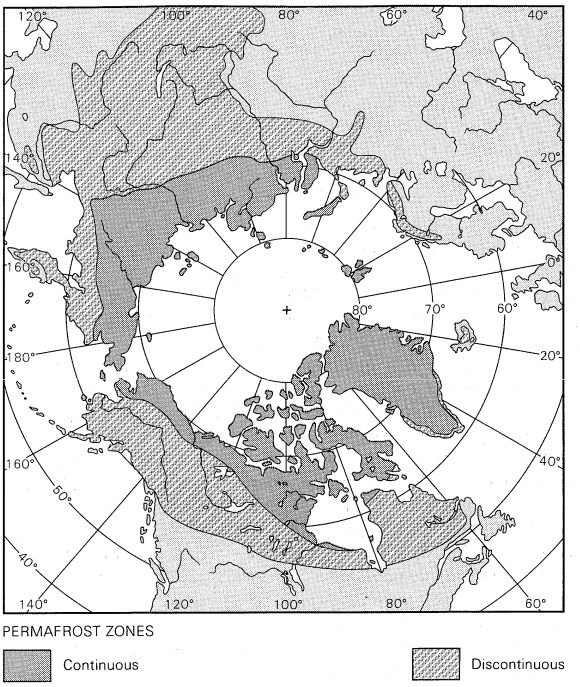


Figure 11: Distribution of permafrost in the Northern Hemisphere.

From T. Pewe, *The Periglacial Environment; Past and Present* (1969); McGill-Queens University Press

Permafrost forms and exists in a climate where the mean annual air temperature is 0° C (32° F) or colder. Such a climate is generally characterized by long, cold winters with little snow and short, relatively dry, cool summers. Permafrost, therefore, is widespread in the Arctic, sub-Arctic, and Antarctica. There is a fairly good correlation between the southern boundary of Northern Hemisphere permafrost and the position of the 0° C mean annual air isotherm.

Although the existence of permafrost had been known to the inhabitants of Siberia for centuries, scientists of the Western world did not take seriously the isolated reports of a great thickness of frozen ground existing under northern forest and grasslands until 1836. Then, Alexander Theodor von Middendorff measured temperatures to depths of approximately 107 metres (350 feet) of permafrost in the Shargin shaft, an unsuccessful well dug for the governor of the Russian-Alaskan Trading Company, at Yakutsk, and estimated that the permafrost was 215 metres (700 feet) thick.

Preservation of animal carcasses

A most interesting aspect of permafrost is the preservation of carcasses and partial carcasses of modern and extinct Ice Age mammals within it. Frozen carcasses have been known in the north for centuries. The first direct report of a carcass of a mammoth in frozen ground of Siberia was made by E. Yssbrants Ides in 1692. At least one or two almost complete, frozen mammoths have been reported from Siberia. Although no such famous frozen carcasses as those found in Siberia have been reported from North America, partial carcasses of mammoth, bison, musk ox, moose, horse, lynx, caribou, and ground squirrel have been found preserved in the permafrost areas of Alaska.

ORIGIN AND STABILITY OF PERMAFROST

In areas where the mean annual air temperature becomes colder than 0° C, some of the ground frozen in the winter will not be completely thawed in the summer; therefore, a layer of permafrost will form and continue to grow downward gradually each year from the seasonally frozen ground. The permafrost layer will become thicker each winter, its thickness controlled by the thermal balance between the heat flow from the Earth's interior and that flowing outward into the atmosphere. This balance depends upon the mean annual air temperature and the geothermal gradient. The average geothermal gradient is an increase of 1° C (1.8° F) for every 30 to 60 metres (100 to 200 feet) of depth. Eventually the thickening permafrost layer reaches an equilibrium depth at which the amount of geothermal heat reaching the permafrost is on the average equal to that lost to the atmosphere. Thousands of years are required to attain a state of equilibrium where permafrost is hundreds of feet thick.

Geothermal gradient

An example of the change of temperature of frozen ground with depth and the upper and lower limit of permafrost is illustrated in Figure 13. The annual fluctuationof air temperature from winter to summer is reflected in a subduedmanner in the upper few feet of the ground. This fluctuation diminishes rapidly with depth, being only a few degrees at 7.5 metres (25 feet), and is barely detectable at 15 metres (50 feet). The level of zero amplitude, at which fluctuations are hardly detectable, is nine to 15 metres (30 to 50 feet). If the permafrost is in thermal equilibrium, the temperature at the level of zero amplitude is generally regarded as the minimum temperature of the permafrost. Below this depth the temperature increases steadily under the influence of heat from the Earth's interior. The temperature of permafrost at the level of zero amplitude is close to the average temperature of the ground surface, being, however, 2°–6° C (3.6°–10.8° F) warmer than the mean annual air temperature determined from the U.S. Weather Bureau-type surface installations (at a height of 1.5 metres [five feet] above ground level). The temperature of permafrost at the depth of minimum annual seasonal change varies from near 0° C at the southern limit of permafrost to −10° C (14° F) in northern Alaska and −13° C (9° F) in northeastern Siberia.

Climatic controls

As the climate becomes colder or warmer, but maintaining a mean annual temperature colder than 0° C, the temperature of the permafrost correspondingly rises or declines, resulting in changes in the position of the base of permafrost. The position of the top of permafrost will be lowered by thawing when the climate warms to a mean annual air temperature warmer than 0° C. The rate at which the base or top of permafrost is changed depends not only on the amount of climatic fluctuation but also on the amount of ice in the ground and the composition of the ground, conditions that in part control the geothermal gradient. If the geothermal gradient is known and if the surface temperature is stable for a long period of time, it is, therefore, possible to predict from a knowledge of the

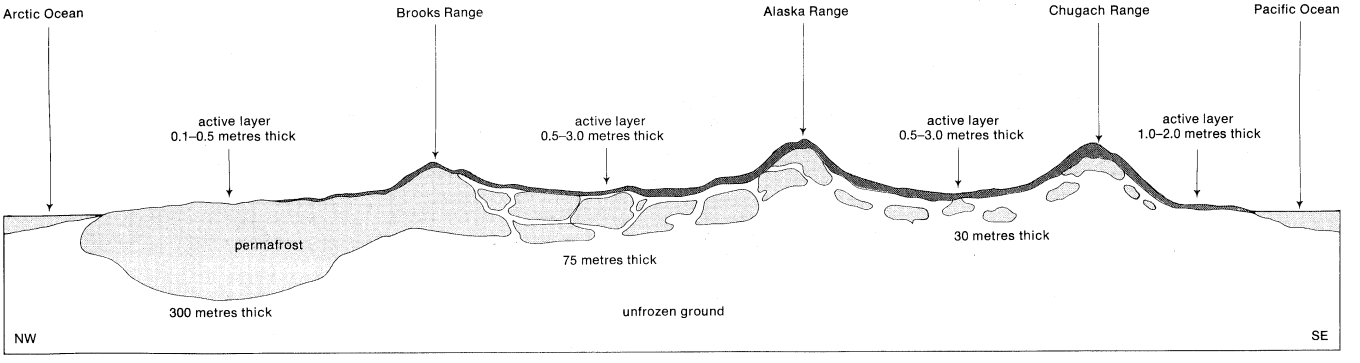


Figure 12: Approximate distribution of permafrost in Alaska.

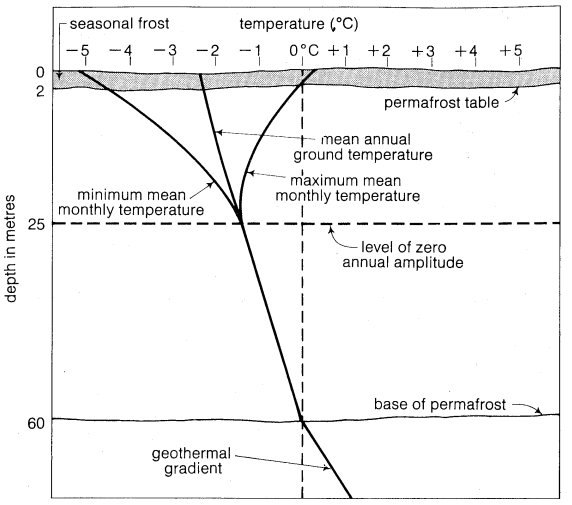


Figure 13: Hypothetical example of a temperature profile and thickness of permafrost in central Alaska.

mean annual air temperature the depth of permafrost in a particular area remote from water bodies.

Permafrost is the result of present climate. Many temperature profiles show, however, that permafrost is not in equilibrium with present climate at the sites of measurement. Some areas show, for example, that climatic warming since the last third of the 19th century has caused a warming of the permafrost to a depth of a few hundred feet. Here some of the existing perennially frozen ground would disappear with rising base of permafrost if present surface conditions were to persist for several thousands of years. Actually, in such areas much of the permafrost is a product of a colder, former climate.

It is thought that permafrost first occurred in conjunction with the onset of glacial conditions near the beginning of the Quaternary Period about 2,500,000 to 3,000,000 years ago. The oldest permafrost known in North America, however, is about mid-Quaternary in age, or less than 2,000,000 years old. In the sub-Arctic at least, most permafrost probably disappeared during the interglacial times and reappeared in glacial times. Most existing permafrost probably formed in the cold (glacial) period of the last 70,000 years.

LOCAL THICKNESS AND ICE CONTENT

The thickness and areal distribution of permafrost is directly affected by snow and vegetation cover, topography, bodies of water, the interior heat of the Earth, and the temperature of the atmosphere as mentioned earlier.

The most conspicuous change in thickness of permafrost is related to climate. At Barrow, Alaska, the mean annual air temperature is −12° C (10° F) and the thickness is 400 metres (1,300 feet). At Fairbanks, Alaska, in the discontinuous zone of permafrost in central Alaska, the mean annual air temperature is −3° C (27° F) and the thickness is about 90 metres (300 feet). Near the southern border of permafrost, the mean annual air temperature is about 0° or −1° C and the perennially frozen ground is only a few feet thick.

If the mean annual air temperature is the same in two areas, the permafrost will be thicker where the conductivity of the ground is higher and the geothermal gradient is less. A.H. Lachenbruch of the U.S. Geological Survey reports an interesting example from northern Alaska. The mean annual air temperatures at Cape Simpson and Prudhoe Bay are similar, but permafrost thickness is 300 metres (1,000 feet) at Cape Simpson and about 600 metres (2,000 feet) at Prudhoe Bay because rocks at Prudhoe Bay are more siliceous and have a higher conductivity and a lower geothermal gradient than rocks at Cape Simpson.

Effect of water bodies

Bodies of water, lakes, rivers, and the sea have a profound effect on the distribution of permafrost. A deep lake that does not freeze to the bottom during the winter will be underlain by a zone of thawed material. If the minimum horizontal dimension of the deep lake is about twice as much as the thickness of permafrost nearby, there probably exists an unfrozen vertical zone extending all the way to the bottom of permafrost. Such thawed areas extending all the way through permafrost are widespread under rivers and sites of recent rivers in the discontinuous zone of permafrost and under major, deep rivers in the far north. Under the wide floodplains of rivers in the sub-Arctic, the permafrost is sporadically distributed both laterally and vertically. Small, shallow lakes that freeze to the bottom each winter are underlain by a zone of thawed material, but the thawed zone does not completely penetrate permafrost except near the southern border of permafrost.

Inasmuch as south-facing hillslopes receive more incoming solar energy per unit area than other slopes, they are warmer; permafrost is generally absent on these in the discontinuous zone (Figure 14) and is thinner in the continuous zone.

The main role of vegetation in permafrost areas is to shield perennially frozen ground from solar energy. Vegetation is an excellent insulating medium and removal or disturbance of it, either by natural processes or by man, causes thawing of the underlying permafrost. In the continuous zone the permafrost table may merely be lowered by the disturbance of vegetation, but in a discontinuous zone permafrost may be completely destroyed in certain areas.

Snow cover also influences heat flow between the ground and the atmosphere and therefore affects the distribution of permafrost. If the net effect of timely snowfalls is to prevent heat from leaving the ground in the cold winter, permafrost becomes warmer. Actually, local differences in vegetation and snowfall in areas of thin and warm permafrost are critical for the formation and existence of the perennially frozen ground. Permafrost is not present in areas of the world where great snow thicknesses persist throughout most of the winter.

Types of ground ice. The ice content of permafrost is probably the most important feature of permafrost affecting human life in the north. Ice in the perennially frozen ground exists in various sizes and shapes and has definite distribution characteristics. The forms of ground ice can be grouped into five main types: (1) pore ice; (2) segregated, or Taber, ice; (3) foliated, or wedge, ice; (4) pingo ice; and (5) buried ice.

1. Pore ice, which fills or partially fills pore spaces in the ground, is formed by pore water freezing *in situ* with no addition of water. The ground contains no more water in the solid state than the ground could hold if the water were in liquid state.

From T. Pewe, *Permafrost and Its Effect on Life in the North* (1966); Oregon State University Press

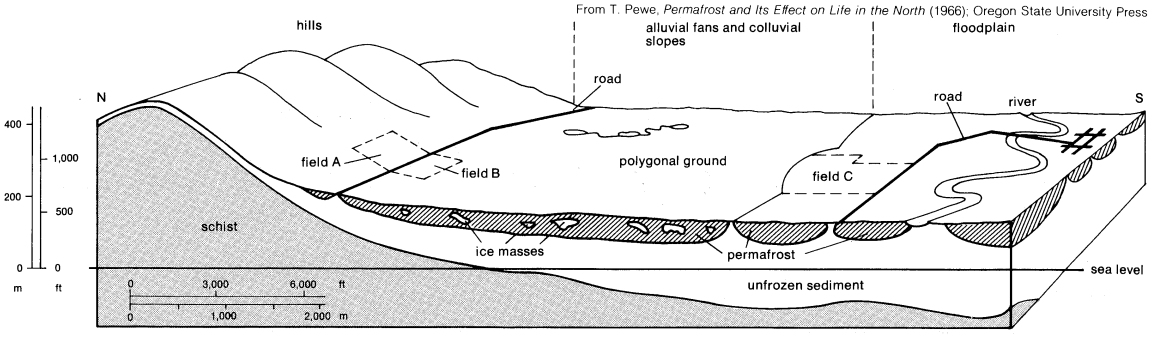


Figure 14: Character and distribution of permafrost in Fairbanks, Alaska, area.

2. Segregated, or Taber, ice includes ice films, seams, lenses, pods, or layers generally 0.15 to 13 centimetres (0.06 to five inches) thick that grow in the ground by drawing in water as the ground freezes. Small ice segregations are the least spectacular but one of the most extensive types of ground ice, and engineers and geologists interested in ice growth and its effect on engineering structures have studied them considerably. Such observers generally accept the principle of bringing water to a growing ice crystal, but they do not completely agree as to the mechanics of the processes. Pore ice and Taber ice occur both in seasonally frozen ground and in permafrost.

3. Foliated ground ice, or wedge ice, is the term for large masses of ice growing in thermal contraction cracks in permafrost.

4. Pingo ice is clear, or relatively clear, and occurs in permafrost more or less horizontally or in lens-shaped masses. Such ice originates from groundwater under hydrostatic pressure.

5. Buried ice in permafrost includes buried sea, lake, and river ice and recrystallized snow, as well as buried blocks of glacier ice in permafrost climate.

Ice content

World estimates of the amount of ice in permafrost vary from 200,000 to 500,000 cubic kilometres (49,000 to 122,000 cubic miles), or less than 1 percent of the total volume of the Earth. Based on an examination of ice in the ground in many bore holes near Barrow, Alaska, and extrapolation of this bore hole information to the rest of the Coastal Plain by use of aerial photographs and geologic maps, it has been estimated that 10 percent by volume of the upper three metres (10 feet) of permafrost on the Coastal Plain of Alaska is composed of foliated ground ice (ice wedges). Taber ice is the most extensive type of ground ice, and in places it represents 75 percent of the ground by volume. It is calculated that the pore and Taber ice content in the depth between 0.5 and three metres (1.5 and 10 feet; surface to 0.5 metre is seasonally thawed) is 61 percent by volume, and between three and nine metres (10 and 30 feet) it is 41 percent. The total amount of pingo ice is less than 0.1 percent of the permafrost. The total ice content in the permafrost of the Arctic Coastal Plain of Alaska is estimated to be 1,500 cubic kilometres (365 cubic miles), and below nine metres (30 feet) most of that is present as pore ice.

Troy L. Pewe, Arizona State University

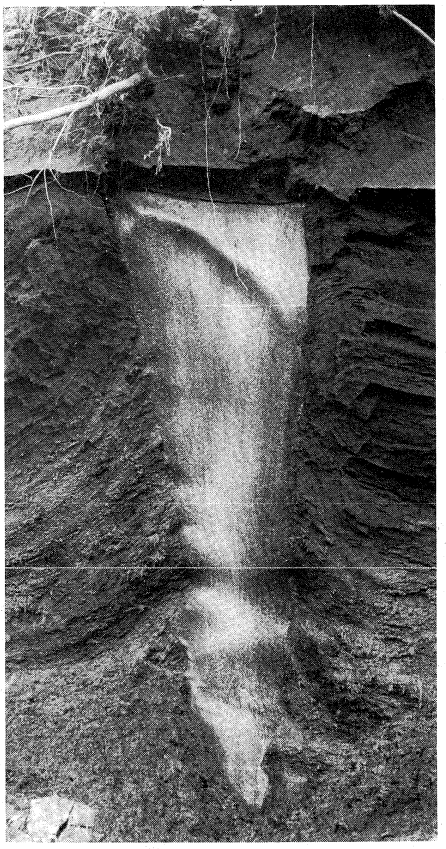


Figure 15: Foliated ground-ice mass (ice wedge) in organic rich silt exposed by gold-mining operations on Wilber Creek, near Livengood, Alaska.

Ice wedges. The most conspicuous and controversial type of ground ice in permafrost is that formed in large ice wedges or masses with parallel or subparallel foliation structures. Most foliated ice masses occur as wedge-shaped, vertical, or inclined sheets or dikes 2.5 centimetres to three metres (one inch to 10 feet) wide and 0.3 to nine metres (one to 30 feet) high when viewed in transverse cross section (Figure 15). Some masses seen on the face of frozen cliffs may appear as horizontal bodies a few inches to three metres (10 feet) in thickness and 0.3 to 14 metres (one to 45 feet) long, but the true shape of these ice wedges can be seen only in three dimensions. Ice wedges are parts of polygonal networks of ice enclosing cells of frozen ground three to 30 metres (10 to 100 feet) or more in diameter.

Thermal contraction theory

The origin of ground ice was first studied in Siberia, and discussions in print of the origin of large ground-ice masses in perennially frozen ground of North America have gone on since Otto von Kotzebue recorded ground ice in 1816 at a spot now called Elephant's Point in Eschscholtz Bay of Seward Peninsula. The theory for the origin of ice wedges now generally accepted is the thermal contraction theory that, during the cold winter, polygonal thermal contraction cracks, about an inch wide and several feet deep, form in the frozen ground; then when, in early spring, water from the melting snow runs down these tension cracks and freezes, a vertical vein of ice is produced that penetrates into permafrost; when the permafrost warms and re-expands during the following summer, horizontal compression produces upturning of the frozen sediment by plastic deformation; then during the next winter, renewed thermal tension reopens the vertical ice-cemented crack, which may be a zone of weakness; another increment of ice is added in the spring when meltwater again enters and freezes. Over the years the vertical wedge-shaped mass of ice is produced (Figure 15).

Ice wedges may be classified as active, inactive, and ice-wedge casts. Active ice wedges are those that are actively growing. The wedge may not crack every year, but during many or most years cracking does occur, and an increment of ice is added. Ice wedges require a much more rigorous climate to grow than does permafrost. The permafrost table must be chilled to −15° C to −20° C (5° to −4° F) for contraction cracks to form. On the average, it is assumed that ice wedges generally grow in a climate where the mean annual air temperature is −7° or −8° C (19° to 18° F) or colder.

The area of active ice wedges appears to roughly coincide with the continuous permafrost zone. From north to south across the permafrost area in North America, a decreasing number of wedges crack frequently. The line dividing zones of active and inactive ice wedges is arbitrarily placed at the position where it is thought most wedges do not frequently crack.

Inactive ice wedges are those that are no longer growing. The wedge does not crack in winter and, therefore, no new ice is added. A gradation between active ice wedges and inactive ice wedges occurs in those wedges that crack rarely. Inactive ice wedges have no ice seam or crack extending from the wedge upward to the surface in the spring. The wedge top may be flat (Figure 15), especially if thawing has lowered the upper surface of the wedge at some time in the past.

Indication of former temperature

Ice wedges in the world are of several ages, but none appear older than the outset of the last major cold period, about 70,000 years ago. Wedges dated by radiocarbon analyses range from 3,000 to 32,000 years in age.

In many places in the now temperate latitudes of the world, in areas of past permafrost, ice wedges have melted, and resulting voids have been filled with sediments collapsing from above and the sides. These ice-wedge casts, or fossil ice wedges, are important as paleoclimatic indicators and indicate a climate of the past with at least a mean annual air temperature of −7° or −8° C or colder.

SURFACE MANIFESTATIONS OF PERMAFROST AND SEASONALLY FROZEN GROUND

Many distinctive surface manifestations of permafrost exist in the Arctic and sub-Arctic, including such geomorphic features as polygonal ground, thermokarst phenomena, and pingos. In addition to the above, there are many features caused in large part by frost action that are common in but not restricted to permafrost areas, such as solifluction (soil flowage) and frost-sorted patterned ground.

Areas underlain by permafrost. One of the most widespread geomorphic features associated with permafrost is the microrelief pattern on the surface of the ground generally called polygonal ground, or tundra polygons (Figure 16). This pattern, which covers thousands of square miles of the Arctic and less in the sub-Arctic, is caused by an intersecting network of shallow troughs (delineating polygons three to 30 metres (10 to 100 feet) in diameter. The troughs are underlain by more or less vertical ice wedges (Figure 15) 0.6 to three metres (two to 10 feet) across on the top that are joined together in a honeycomb network. These large-scale polygons should not be confused with the small-scale polygons or patterned ground produced by frost sorting.

Ice-wedge polygons

The ice-wedge polygons may be low centred or high centred. Upturning of strata adjacent to the ice wedge may make a ridge of ground on the surface on each side of the wedge, thus enclosing the polygons. Such polygons are lower in the centre and are called low-centre polygons or raised-edge polygons and may contain a pond in the centre. Low-centre, or raised-edge, polygons indicate that ice wedges are actually growing and that the sediments are being actively upturned. If erosion, deposition, or thawing is more prevalent than the up-pushing of the sediments along the side of the wedge or if the material being pushed up cannot maintain itself in a low ridge, the low ridges will be absent, and there may be either no polygons at the surface or the polygons may be higher in the centre than the troughs over the ice wedges that enclose them. Both high-centre and low-centre tundra polygons are widespread in the polar areas and are good indicators of the presence of foliated ice masses; care must be taken, however, to demonstrate that the pattern is not a relic and an indication of "fossil ice" wedges.

In many parts of the temperate latitudes of Asia, Europe, and North America, incompletely developed or poorly developed polygonal ground occurs on the same scale as in the Arctic. These large-scale polygons in the nonpermafrost areas are excellent evidence of the former extent of permafrost and ice wedges in the last glacial period.

In many areas of the continuous permafrost zone surface, drainage follows the troughs of the polygons (tops of the ice wedges); and at ice wedge junctions, or elsewhere, melting may occur to form small pools. The joining of these small pools by a stream causes the pools to resemble beads on a string, a type of stream form called beaded drainage. Such drainage indicates the presence of perennially frozen, fine-grained sediments cut by ice wedges.

Thermokarst topography and pingos

The thawing of permafrost creates thermokarst topography, an uneven surface that contains mounds, sinkholes, tunnels, caverns, and steep-walled ravines caused by melting of ground ice. The hummocky ground surface resembles karst topography found in limestone areas. Thawing may result from artificial or natural removal of the vegetation or from warming of the climate (Figure 17).

Thawed depressions filled with water (thaw lakes, thermokarst lakes, cave-in lakes) are widespread in permafrost areas, especially in those underlain with perennially frozen silt. They may occur on hillsides or even on hilltops and are good indicators of ice-rich permafrost. Locally, deep thermokarst pits six metres (20 feet) deep and nine metres (30 feet) across may form as ground ice melts. These openings may exist as undetected caverns for many years before the roof collapses. Such collapses in agricultural or construction areas are real dangers. Thermokarst mounds are polygonal or circular hummocks three to 15 metres (10 to 50 feet) in diameter and 0.3 to 2.5 (one to eight feet) high that are formed as a polygonal network of ice melts and leaves the inner-ice areas as mounds.

The most spectacular landforms associated with permafrost are pingos, small ice-cored circular or elliptical hills of frozen sediments or even bedrock, three to more than 60 metres (10 to 200 feet) high and 15 to 450 metres (50 to 1,500 feet) in diameter. Pingos are widespread in the continuous permafrost zone and are quite conspicuous because they rise above the tundra. They are much less conspicuous in the forested area of the discontinuous permafrost zone. They are generally cracked on top with summit craters formed by melting ice. There are two types of pingos, based on origin. The closed-system type forms in level areas when unfrozen groundwater in a thawed zone becomes confined on all sides by permafrost, freezes, and heaves the frozen overburden to form a mound. This type is larger and occurs mainly in tundra areas of continuous permafrost. The open-system type is generally smaller and forms on slopes when water beneath or within the permafrost penetrates the permafrost under hydrostatic pressure. A hydrolaccolith (water mound) forms and freezes, heaving the overlying frozen and unfrozen ground to produce a mound.

Present pingos are apparently the result of postglacial climate and are less than 4,000–7,000 years old. Pingos were present in now temperate latitudes during the latest glacial epoch and are now represented by low circular ridges enclosing bogs or lowlands.

Troy L. Pewe, Arizona State University

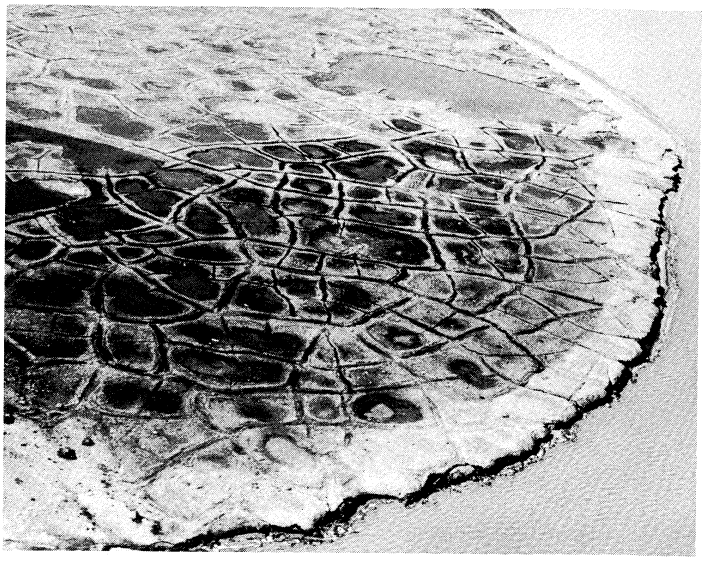


Figure 16: Raised-edge ice wedge polygons on the seacoast near Barrow, Alaska, in summer. The polygons are from seven to 15 metres in diameter.
By courtesy of the U.S. Army (CRREL); photograph, Robert Lewellen

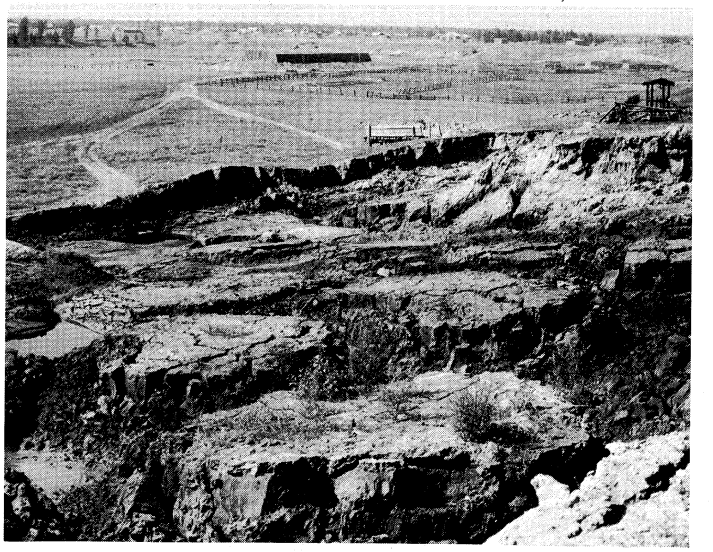


Figure 17: Modern thermokarst formation at Maya, near Yakutsk, Siberia.

Near the southern border of permafrost occur palsas, low hills and knobs of perennially frozen peat about 1.5 to six metres (five to 20 feet) high, evidently forming with accumulation of peat and segregation of ice.

Features related to seasonal frost. Many microgeomorphic features common to the periglacial environment may or may not be associated with permafrost. Intense seasonal frost action, repeated freezing and thawing throughout the year, produces small-scale patterned ground. Repetitive freezing and thawing tends to stir and sort granular sediments, thus forming circles, stone nets, and polygons a few centimetres to six metres (20 feet) in diameter. The coarse cobbles and boulders form the outside of the ring and the finer sediments occur in the centre. The features require a rigorous climate with some fine-grained sediments and soil moisture, but they do not necessarily need underlying permafrost. Permafrost, however, forms an impermeable substratum that keeps the soil moisture available for frost action. On gentle slopes the stone nets may be distorted into garlands by downslope movement or, if the slope is steep, into stone stripes 0.3 to 0.6 metre (one to two feet) wide and 30 metres (100 feet) long.

Stone stripes, stone nets, patterned ground, and solifluction

In areas underlain by an impermeable layer (seasonally frozen ground or perennially frozen ground), the active layer is often saturated with moisture and is quite mobile. The progressive downslope movement of saturated detrital material under the action of gravity and working in conjunction with frost action is called solifluction. This material moves in a semifluid condition and is manifested by lobelike and sheetlike flows of soil on slopes. The lobes are up to 30 metres (100 feet) wide and have a steep front 0.3 to 1.5 metres (one to five feet) high. An outstanding feature of solifluction is the mass transport of material over low-angle slopes. Solifluction deposits are widespread in polar areas and consist of a blanket 0.3 to 1.8 metres (one to six feet) thick of unstratified or poorly stratified, unsorted, heterogeneous, till-like detrital material of local origin. In many areas the terrain is characterized by relatively smooth, round hills and slopes with well-defined to poorly defined solifluction lobes or terraces. If the debris is blocky and angular and fine material is absent, the lobes are poorly developed or absent. Areas in which solifluction lobes are well formed lie almost entirely above or beyond the forest limit.

In many areas the frost-rived debris contains few fine materials, little water, and consists of angular fragments of well-jointed, resistant rock. Under such circumstances, solifluction lobes do not often occur, but instead striking sheets or streams of angular rubble form. These are called rock streams or rubble sheets.

PROBLEMS POSED BY PERMAFROST

Development of the north demands an understanding of and the ability to cope with problems of the environment dictated by permafrost. Although the frozen ground hinders agricultural and mining activities, the most dramatic, widespread, and economically important examples of the influence of permafrost on life in the north involve construction and maintenance of roads, railroads, airfields, bridges, buildings, dams, sewers, and communication lines. Engineering problems are of four fundamental types: (1) those involving thawing of ice-rich permafrost and subsequent subsidence of the surface under unheated structures such as roads and airfields; (2) those involving subsidence under heated structures; (3) those resulting from frost action, generally intensified by poor drainage caused by permafrost; and (4) those involved only with the temperature of permafrost that causes buried sewer, water, and oil lines to freeze.

A thorough study of the frozen ground should be part of the planning of any engineering project in the north. It is generally best to attempt to disturb the permafrost as little as possible in order to maintain a stable foundation for engineering structures, unless the permafrost is thin; then, it may be possible to destroy the permafrost. The method of construction preserving the permafrost has been called the passive method; alternately, the destroying of permafrost is the active method.

Because thawing of permafrost and frost action are involved in almost all engineering problems in polar areas, it is advisable to consider these phenomena generally. The delicate thermal equilibrium of permafrost is disrupted when the vegetation, snow cover, or the active layer is compacted. The permafrost table is lowered, the active layer is thickened, and considerable ice is melted. This process lowers the surface and provides (in summer) a wetter active layer with less bearing strength. Such disturbance permits a greater penetration of summer warming. It is common procedure to place a fill, or pad, of gravel under engineering works. Such a fill generally is a good conductor of heat and, if thin, may cause additional thawing of permafrost. The fill must be made thick enough to contain the entire amplitude of seasonal temperature variation—in other words, thick enough to restrict the annual seasonal freezing and thawing to the fill and the compacted active layer. Under these conditions no permafrost will thaw. Such a procedure is quite feasible in the Arctic, but in the warmer sub-Arctic it is impractical because of the enormous amounts of fill needed. Under a heated building, profound thawing may occur more rapidly than under roads and airfields.

Frost action, the freezing and thawing of moisture in the ground, has long been known to seriously disrupt and destroy structures in both polar and temperate latitudes. In the winter the freezing of ground moisture produces upward displacement of the ground (frost heaving), and in the summer excessive moisture in the ground brought in during the freezing operation causes loss of bearing strength. The intensity of frost action is regulated by temperature, the texture of the ground, and the amount of soil moisture. Frost action is best developed in silt-sized and silty clay-sized sediments in areas of rigorous climate and poor drainage. Polar latitudes are ideal for maximum frost action because most lowland areas are covered by fine-grained sediments, and the underlying permafrost causes poor drainage.

Frost action and man-made structures

Piles are used to support many, if not most, structures built on ice-rich permafrost. In regions of cold winters, many pile foundations are in ground subject to seasonal freezing and, therefore, possibly subject to the damaging effect of frost heaving, which tends to displace the pile upward and thus to disturb the foundation of the structure. The displacement of piling is not limited to the far north, though maximum disturbance probably is encountered most widely in the sub-Arctic. Expensive maintenance and sometimes complete destruction of bridges, school buildings, military installations, and other structures have resulted from failure to understand the principles of frost heaving of piling.

The need for careful geological location of highways, railroads, airfields, and pipelines is very important anywhere in the world; but it is absolutely essential in permafrost areas in order to avoid constant, costly maintenance—if not complete abandonment. Such location studies were not frequently considered in construction work before the 1950s in North America and perhaps before the 1930s in Asia. Even late in the 20th century, many construction projects proceeded with inadequate location studies.

Highways in polar areas are relatively few and mainly unpaved. They are subject to subsidence by thawing of permafrost in summer, frost heaving in winter, and loss of bearing strength on fine-grained sediments in summer. Constant grading of gravel roads permits maintenance of a relatively smooth highway. Where the road is paved over ice-rich permafrost, the roadway becomes rough and is much more costly to maintain than are unpaved roads. Many of the paved roads in polar areas have required resurfacing two or three times in a 10-year period.

Railroads particularly have serious construction problems and require costly upkeep in permafrost areas because of the necessity of maintaining a relatively low gradient and the subsequent location of the roadbed in ice-rich lowlands that are underlain with perennially frozen ground. The Trans-Siberian Railroad, the Alaska Railroad, and some Canadian railroads in the north are locally underlain by permafrost with considerable ground ice. As the large masses of ice melt each summer, constant maintenance is required to level these tracks. In winter, extensive mainte-

nance is also required to combat frost heaving when local displacements of 2.5 to 35 centimetres (one to 14 inches) occur in roadbeds and bridges.

Heated buildings in polar areas can have a conventional basement if they are built in areas of no permafrost or of bedrock and gravel with little or no ice. Construction of such buildings on ice-rich permafrost, however, requires a technique that enables the cold winter air to circulate under the building to counteract the heat generated in the building. The most common way to build structures above the ground with large spaces under the foundation for airflow is to construct on piles, a technique perfected both in the U.S.S.R. and in North America. The piles must be erected so that they do not move. If the piles are jacked up in the winter by frost action, the building will be disturbed and can be destroyed. It is therefore important that the pile be adequately anchored in permafrost or it be isolated from the heaving of the ground by a movable collar.

Groundwater and agriculture

Both surface and subsurface water supplies in the north are beset by problems uncommon in temperate latitudes. Groundwater sources are above the permafrost, in unfrozen zones within the permafrost, and below the permafrost. Sources above permafrost may be seasonal and subject to contamination from cesspools. Sources from unfrozen zones within the permafrost are common in the discontinuous zones where water also can be obtained below permafrost. In the far north, however, little water is available in the permafrost, deep drilling 300–600 metres [1,000–2,000 feet] is expensive, and the sediments generally lack sufficient groundwater.

Distribution of water and sewage presents unique problems in permafrost areas. In addition to requiring a heated distribution system to prevent freezing in permafrost or seasonal frost areas, utility distribution lines are subject to problems because of frost action and thawing of permafrost. The most successful construction method generally used in such situations is to place all services in heated distribution boxes called utilidors, which enclose water, steam, gas, sewage, and other utilities. The utilidor may be either under or above ground.

Permafrost affects agricultural developments in many parts of the discontinuous permafrost zone. Its destructive effect on cultivated fields in both the U.S.S.R. and North America results from the thawing of large masses of ice in the permafrost. If care is not exercised in selecting areas to be cleared for cultivation, thawing of the permafrost may necessitate abandonment of fields or their reduction to pasturage. As illustrated in Figure 14, Field A, on the southward-facing hillside, is not affected by permafrost; Field B, on the alluvial fan, however, is underlain with permafrost containing large masses of ground ice. When Field B is cleared, thermokarst topography will form. Field C, on the floodplain, is underlain with permafrost without large ice masses, and so no thermokarst pits and mounds will form when this ground is cleared.

In summary, it can be said that in construction and maintenance of railroads, buildings, water and sewer lines, dams, roads, bridges, and airfields, normal techniques must often be modified at additional costs because of permafrost. But, despite special problems unique to polar areas, development of the permafrost regions will continue at an ever-increasing rate. Man has already learned to cope with the problems, and improvements of scientific and engineering approaches, with careful geological site selection and study of the permafrost problem, will allow successful expansion into polar areas.

(T.L.Pe./Ed.)

Icebergs and pack ice

Ice in the waters of the Earth's polar regions occurs in two forms, namely, pack ice and icebergs. Pack ice forms from seawater and is generally only one to two years old, whereas icebergs are fragments of ice sheets and glaciers that formed on land areas during intervals of thousands of years. Pack ice expands during winter to cover large areas of the oceans in both hemispheres. Melting occurs in spring and summer and the margins of the pack ice retreat. This warmer weather aids calving (separation) of icebergs at the boundaries of ice sheets and glaciers, however, and many icebergs start their transit toward the Equator at this time.

Table 1: World Distribution of Icebergs and Pack Ice*

	area ($\times 10^6$)		volume ($\times 10^6$)		thickness	
	sq mi	sq km	cu mi	cu km	ft	m
Antarctic Ice Sheet	5.2	13.5	5.8	24.1	7,200	2,195
Greenland Ice Sheet	.67	1.7	.62	2.58	4,900	1,494
Greenland glaciers	.023	.060	.012	.050		
Antarctic pack ice						
Winter	7.7	19.9				
Summer	1.5	3.9	5.5×10^{-3}	22.9×10^{-3}	6–10	1.8–3.0
Arctic pack ice						
Winter	4.6	11.9				
Summer	3.5	9.1	2.2×10^{-3}	9.2×10^{-3}	10–13	3–4
Western North Atlantic annual melting	.467	1.210				
Arctic icebergs	.003	.008	1.82×10^{-3}	7.57×10^{-3}		

*The area and volume of the world ocean, from which the ice is principally derived, are 361 $\times 10^6$ square kilometres and 1,370 $\times 10^6$ cubic kilometres.

Importance of at sea

This ice at sea is of substantial importance to man. Major shipping routes around the world could be shortened by as much as 30 percent if regular navigation through the Arctic Ocean became feasible, for example. The economics of freshwater extraction from icebergs also is instructive. An iceberg 16 kilometres (10 miles) long could be hauled into the Peru (Humboldt) Current, which flows from the south along South America, and ultimately to Los Angeles where it could supply fresh water to that city for three weeks. Factors of wind effect, breakup, grounding, accessible supply, and controlled melting on delivery argue against this venture, however. In the Northern Hemisphere about 10,000 icebergs are produced each year from the West Greenland glaciers and an average of 375 flow south of Newfoundland into the North Atlantic shipping lanes.

Because ice reflects four to five times more sunlight than does the ocean, the change in areal coverage of sea ice between winter and summer is climatically important. There is a change in areal coverage of 20 percent from winter to summer in Arctic pack ice extent, and 80 percent in Antarctic pack ice maximum to minimum coverage. This winter to summer change in area of pack ice coverage is equivalent to twice the area of the United States.

ICEBERGS

Formation and distribution. Probably the first mention of icebergs was that of St. Brendan, an Irish monk whose partly fictional writings suggest that he encountered a "floating crystal castle" on the high seas. After calving (breaking off) from the Greenland and Antarctic ice sheets and smaller outlying glaciers, icebergs can move thousands of miles in a few years; those in the Arctic can slip down along the eastern North American coast to be caught up in the Gulf Stream and, while melting, be carried in a few weeks to within a few hundred miles of England and Ireland, or by a combination of wind and current to Bermuda, as happened in 1907 and 1926. These and other rare sightings are shown in Figure 18.

Gulf Stream transport

As previously noted, when snow and freezing rain continue to precipitate over a continent in excess of evaporation, glaciers will form and icebergs will break off their ends and appear in the surrounding ocean. At the point where glaciers or large, extensive ice shelves meet the sea, water pressure beneath the ice shelf or glacier tongue interacts with the outward creeping glacier. The tides, which have ranges up to six metres (20 feet) in the Arctic, along with small sea level changes associated with wind and swells, result in an intermittent increase and decrease in force on the protruding end of the glacier or ice shelf resulting in the birth of a large monolith of drifting ice. There are other ways in which an iceberg is formed. One, which is characteristic of southern Greenland glaciers, consists of a melting or evaporation of the surface portions of the

glacier near its terminus at a greater rate than the water erosion on its underside. This results in an underwater shelf and eventually, through the erosion of water and periodic tidal and other hydraulic forces, this is broken off and an iceberg floats to the surface. Icebergs of varying shapes are produced in this way. A third mechanism by which icebergs are formed is through gradual break off from a hanging glacier or ice shelf. The type of iceberg mechanism is related to the surrounding topography, the climate, and the rate of flow of associated glaciers.

Movement of glaciers producing icebergs

The speed of the ice sheet spread, or creep, over Greenland and the Antarctic varies from zero near their centres to as much as 10 kilometres (six miles) per year in the ice streams that make up glaciers. For the same inclination relative to gravity, ice moves 1/10,000 as fast as water. The average movement in the Antarctic is 360 metres per year (1,200 feet per year), with most of the measurements between extremes of 110 and 1,100 metres per year (360 and 3,600 feet per year). These measurements are obtained from glaciers near the coast and thus are much higher than the average flow rate over the entire Greenland or Antarctic ice sheets. The fastest flowing glaciers have been measured along the western shore of Greenland. The one known as the Quarayaq Glacier flows at a velocity between 20 and 24 metres per day (65 and 80 feet per day), greater than the velocity of most alpine glaciers. Jakobshavn Glacier at latitude 70° N, approximately, produces 10 percent of all the Greenland icebergs (approximately 1,350 annually), and flows at about 20 metres (65 feet) per day. The icebergs accumulate in a fjord, and periodically spill from this fjord in groups accompanied by noise that can be heard for several miles. This greatest iceberg-producing glacier measures only seven kilometres (4.4 miles) along its front and is 90 metres (300 feet) above sea level. In contrast to the rapidly moving glaciers, the very wide glaciers that move slowly produce only very small icebergs. Some glaciers, such as the Frederikshåb, Greenland, with a 33-kilometre (21-mile) front, have rates of flow equal to the rate of melting, and thus produce no icebergs. The annual yield of icebergs in the Arctic is, at the most, 15,000, with only 5,000 or so of sufficient size to reach the open ocean intact.

The largest glacier in the Northern Hemisphere is the Petterman Glacier at 81° N 62° W. Although it is only a few yards above seawater, its ice foot extends as far as 40 kilometres (25 miles) out to sea, and it pushes a path through old piled up sea ice. These long fingers of glacier ice, under severe climatic conditions, break off once every 10 to 20 years. Another glacier of importance is the Jungersen Glacier in northern Greenland. This glacier and the Petterman Glacier produce very large tabular icebergs, known in the Arctic as "ice islands." They are similar in shape and mode of formation to the large tabular icebergs of the Antarctic but are much smaller.

East Greenland icebergs tend to move northward; they are small and few in number. The icebergs that do leave

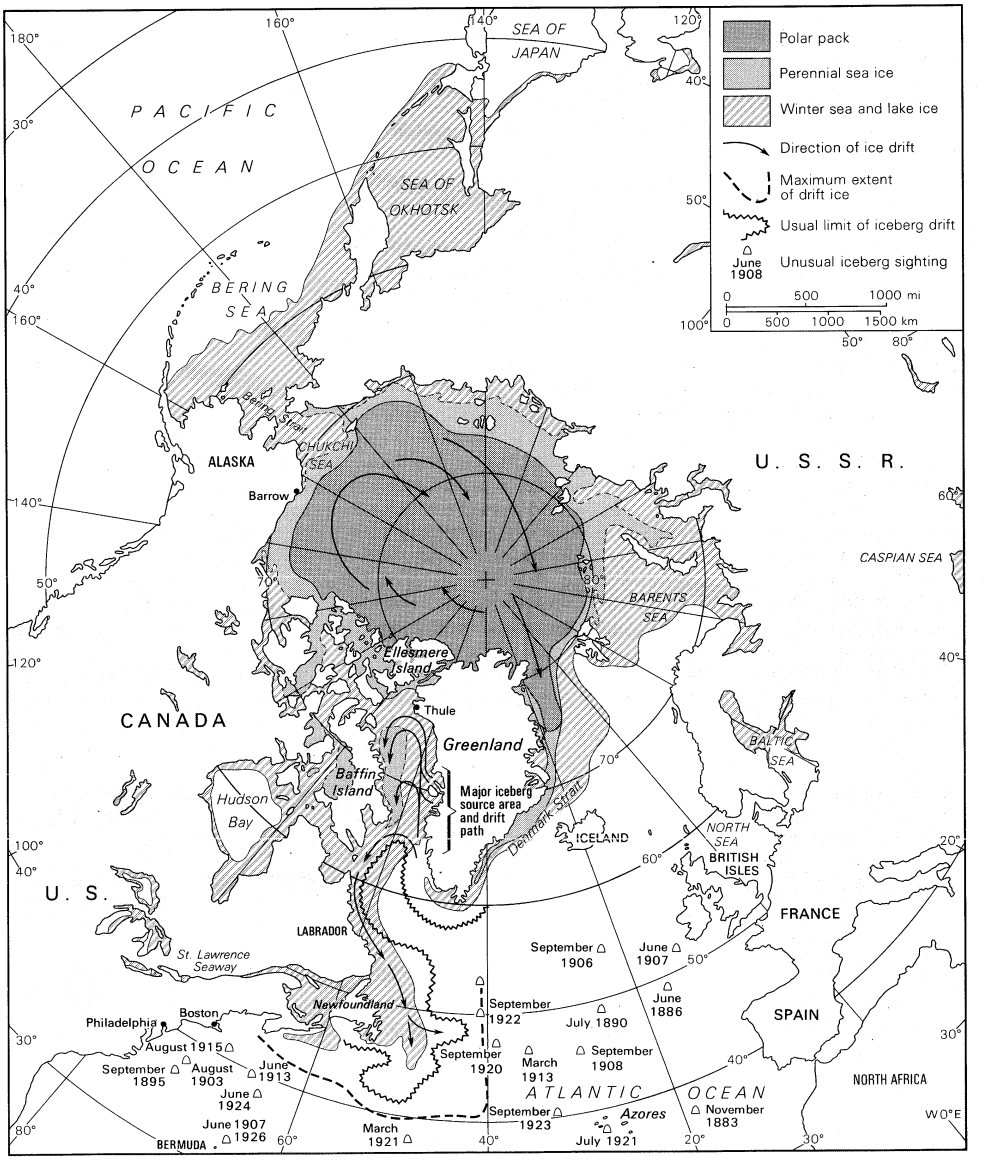


Figure 18: Limits of sea ice and icebergs in the Northern Hemisphere.

the fjord and the coast area enter the East Greenland Current, and some join the West Greenland icebergs. The icebergs that reach the North Atlantic Ocean from northern Greenland, Siberia, and the Northwest Territories constitute less than 10 percent of the total icebergs produced in the Northern Hemisphere. They do not affect the sea routes and, other than their formation of ice islands, have no importance to man. Most icebergs originating from western Greenland are of great importance; of an annual production of about 7,500 bergs (Arctic total is 10,000–15,000), the Labrador Current carries 800 to 1,000 into the open ocean.

Sightings in the North Atlantic

Of the 10,000 to 15,000 icebergs calved from glaciers annually in the Arctic, only 375 on the average pass Newfoundland, or latitude 48° N, into the North Atlantic Ocean. In some years more than 1,000 are seen, in others fewer than 30. The yearly average diminished over 20 years until 1972, when 1,400 icebergs were sighted south of 48° N in the North Atlantic. The distribution of these icebergs in April, May, June, and July is shown in Figure 18.

There are few icebergs in the Arctic Basin proper. The major source area for icebergs in the Barents Sea is Franz Josef Land. Icebergs are not found in the North Pacific except in sounds along the Alaskan-Canadian coast between 55° N and 60° N latitudes.

In the Southern Hemisphere the maximum limit of iceberg drift is 1,600 kilometres (1,000 miles) farther north than the northern extent of sea ice in the Antarctic (Figure 19). Most icebergs are concentrated south of the Antarctic current convergence at about 60° S. Some icebergs from the ice shelves of Antarctica drift north of 42° S latitude in the Atlantic Ocean, but only to latitude 56° S in the South Pacific Ocean. One Antarctic iceberg was sighted only 30 miles south of the Cape of Good Hope (Africa) in 1850. The northern limit of ice sighted in the Southern Hemisphere was 26°30′ S.

Size of icebergs. Arctic icebergs vary in size from the size of a large piano, called growlers, to the dimensions of a 10-story building. Icebergs about the size of a small house are called bergy bits. Many icebergs in the Arctic are about 45 metres (150 feet) tall and 180 metres (600 feet) long. Icebergs of the Antarctic not only are far more abundant but are of enormous dimensions compared to those in the Arctic. Ninety-three percent of the world's mass of icebergs is found surrounding the Antarctic.

Other than Arctic Basin ice islands, the largest iceberg in the Northern Hemisphere was 11 kilometres (seven miles) long and 5.9 kilometres (3.7 miles) wide; it was sighted near Baffin Land in 1882. The largest Arctic iceberg sighted south of Newfoundland was encountered by a convoy of ships during World War II at 43°10′ N and 49°33′ W. There were multiple collisions and much confusion before all ships safely limped into port. This ice island was 1,370 metres (4,500 feet) long, 1,100 metres (3,600 feet) wide, and 18 metres (60 feet) high. These icebergs are similar in size to Antarctic icebergs, but by and large only a few large tabular icebergs are seen in the Arctic as compared to the predominance of the tabular icebergs noted in the Antarctic. The tallest icebergs known were measured at 134 and 158 metres (447 and 527 feet) high; the latter was measured and photographed from a helicopter in the late 1950s.

Antarctic icebergs are characterized by their tremendous size and tabular shape. Lengths up to five miles are not unusual, with ice 45 metres (150 feet) above water. The discovery of the origin of these immense tabular bergs was made in 1841 when the Ross Sea was penetrated and the Ross Ice Shelf was discovered to be afloat. Most Antarctic icebergs are formed from the Antarctic continental ice sheet as it thins toward the coast and exudes into the ocean as a great ice shelf with fronts hundreds of miles long. The four major ice shelves are the Filchner Shelf in the Weddell Sea, the Ross Ice Shelf in the Ross Sea, the Shackleton Ice Shelf in the Indian Ocean sector, and the Larsen Ice Shelf on the Antarctic Peninsula (also known as Palmer Peninsula and Graham Land). Antarctic icebergs, if they remain locked in the pack ice, will last for many years. One of the largest icebergs sighted was over 140 kilometres (90 miles) in length. This tabular iceberg was first sighted in 1927 and presumably the same iceberg was later seen in 1931, at which time it was 100 kilometres (60 miles) long. The largest known Antarctic iceberg was measured by the icebreaker U.S.S. "Glacier" in 1956; it had a length of 333 kilometres (208 miles) and a width of 100 kilometres (60 miles).

Age and melting. Greenland icebergs 300 to 450 metres (1,000 to 1,500 feet) thick represent several centuries of precipitation. This figure is based on the comparison of the thickness of the ice sheet to the known or average annual precipitation of 20 to 60 centimetres (eight to 24 inches) per year in the source area for icebergs. Age of ice in central Greenland, close to bedrock, is estimated at 30,000 to 150,000 years old. The longest core retrieved from the Greenland ice sheet was 1,370 metres (4,580 feet) long with a bottom date of 100,000 years. It is probable that the oldest ice melts before reaching the outlet glaciers. From the known amount of precipitation over Greenland, and the assumption that the ice volume and precipitation rate now are approximately the same as they were many centuries ago, it can be estimated that the mean age of icebergs is 5,000 years. Measurements by carbon-14 dating of entrapped air show that icebergs are hundreds to thousands of years old—the oldest actual measurement being 3,000 years. Arctic ice islands and giant Antarctic bergs last as long as ten years at high latitude. Most icebergs from western Greenland melt within two years of calving from the parent glacier.

Deterioration of icebergs at sea

Once an Arctic iceberg has been calved and moves out to the open sea, it sojourns in Baffin Bay for three months to two years, during which time it undergoes some disintegration through melting and calving of small chunks of ice from its perimeter. This results in a decrease in mass of about 90 percent by the time it reaches the coast of Newfoundland and the Grand Banks in the North Atlantic. When the iceberg enters the region of the Grand Banks, where the warm waters of the Gulf Stream meet the colder waters of the Labrador Current, it has only a few days of life remaining. A large iceberg 120 metres (400 feet) long melted within 36 hours in 80° F water. The estimated rate of iceberg melting is based on the observation of a number of individuals from the International Ice Patrol. For mild sea conditions an iceberg deteriorates at a rate of height decrease of two metres (six feet) per day in 0° C to 4° C (32° F to 40° F) water, and three metres (10 feet) per day in 4° C to 10° C (40° F to 50° F) water. Destruction of icebergs in warm water is increased during stormy weather, when mechanical erosion of icebergs is added to the thermal effects of air and water. During the erosion process icebergs usually take on the form of a saddle, because erosion at one pole of the major axis of the iceberg results in that point rising, while the other end of the major axis is being eroded. Subsequently the latter end, due to loss in weight, arises and this rocking back and forth continues while constant erosion is occurring along the minor axis leading, usually, to a bipeaked or saddle-shaped structure. Table 2 is an estimate of the time necessary for deterioration and is based on unpublished quantitative measurements in 1960 and general observations by ice scientists.

Wind and current effects. Iceberg movement is influenced by direct wind push on its exposed area to an extent far greater than commonly assumed. Although the bulk of the iceberg is below water, in many situations wind has a

Table 2: Representative Deterioration or Melting Times for Icebergs
(at 45° N; estimated)

seawater temperature		melting time (in days)	
°F	°C	*	†
32	0	40	80
40	4	10	20
70	21	4	8

*For an iceberg 80 feet high, 300 feet long.
†For an iceberg greater than 150 feet high, 500 feet long.

dominant influence on the movement. The wind intensity and direction over Baffin Bay in the spring of one year influences the number of icebergs that slip into the North Atlantic that year and the following year. This can be understood by noting that icebergs pouring out of the Arctic north of Newfoundland in spring will run aground or be trapped in the embayments of western Baffin Bay unless wind and current deflect them to the southeast. Of more importance is the effect of wind over western Greenland, however, where intense offshore early summer winds over the ice fjords drive sea ice-entrapped bergs into the West Greenland Current, thus increasing the number of icebergs that, having made the usual counterclockwise circuit, will arrive off Newfoundland the following spring. There is a reasonable correlation between the atmospheric pressure distribution one year and the number of icebergs drifting south of Newfoundland in the following year.

Effect of size and shape

The day to day movement of an iceberg is controlled by the size and shape of the iceberg, previous and present wind, surface wind current, and general ocean current. The most important factor in assessing wind drift of icebergs is size and shape. Although most icebergs have a specific gravity of 0.9, and thus six-sevenths of the mass is below the sea surface, it is not true that this means that a 30-metre- (100-foot-) high iceberg is 180 metres (600 feet) deep in all cases. This is true only for the rectangular, blocky, or flat-topped icebergs common to the Antarctic. Table 3 gives the relation between exposed and underwater areas for various shapes of icebergs.

Table 3: Exposed-to-Submerged Proportions, and Wind Factors, for Icebergs

iceberg description	proportions (exposed : underwater)	wind factor
Flat-topped or blocky	1 : 6	.004
Rounded or domed	1 : 4	.005
Usual Greenland iceberg	1 : 3	.01
Pinnacled or drydock	1 : 2	.03
Winged	1 : 1	.04

Windage and drift

Winged icebergs, those with sail-like pinnacles around the central mass, are very much influenced by the winds and move at speeds of one knot, or 24 nautical miles per day, under the influence of steady winds of 30 knots. The wind force on an iceberg does not result in movement directly downwind, but because of the rotation of the Earth (Coriolis effect), windage on an iceberg is 30 to 50 degrees to the right in the Northern Hemisphere and to the left in the Southern Hemisphere.

The momentum of icebergs is so great that once in motion they continue for hours after the wind has abated. It is possible for an iceberg to be driven before a 30-knot wind at one knot in a direction 30 degrees to the right of the wind, and after the wind stops the iceberg will slow and circle in what is known as an inertial circle, associated with the rotation of the Earth. An iceberg near the Grand Banks will be moving in a direction opposite to that toward which the wind was blowing about eight hours after the wind stops; however, the speed would be low and such anomalous movement would not be observed unless the initial speed and momentum were great. A careful accounting of berg underbody dimensions, and the relationship of wind-current forces and iceberg mass, will lead to explanations of anomalous iceberg movements. To emphasize the importance of wind effect and its prediction, one potentially disastrous case will be cited. In 1960, a 60-metre- (200-foot-) high, 300-metre- (1,000-foot-) long iceberg was driven off the tail of the Grand Banks into the shipping lanes by northwesterly winds averaging 80 kilometres (50 miles) per hour. This iceberg moved 140 kilometres (90 miles) at as much as 3 knots across the Labrador Current and resulted in an emergency move of the North Atlantic shipping lanes to the south.

Iceberg and pack ice drift

The drift of icebergs and pack ice is the result of the ocean current and the wind. When the winds are variable or less than 32 kilometres (20 miles) per hour, and the current greater than 0.5 knot, the current predominates, but when steady 30-knot winds blow for more than 12 hours, the wind effect becomes important even in areas where the ocean current is one to two knots. In addition to windage on the iceberg and the ocean gradient current, the wind induced surface current has the effect of increasing drift speed by about 10 percent for small icebergs and increasing the angle of drift direction.

Table 4: Wind Factor for Speed of Sea-Ice Drift

sea covered by ice (percent)	wind factor	
	rough hummocky surface	smooth floes
10	.08	.01
50	.05	.005
90	.02	.003

Sea ice drift is better understood, but in some respects it is more complicated than iceberg drift. In addition to size, inertia, and exposed-to-underwater dimensions, important additional information on surface roughness, water drag, and internal resistance due to ice-field concentration is needed to describe the motion. The observations of American and Soviet scientists drifting on ice islands in the Arctic Basin, Baffin Bay, and Gulf of St. Lawrence, along with Japanese and Soviet long-term ice-field observations, are summarized in Table 4. As noted from the table, ice fields consisting of 10 percent ice and 90 percent open water will move at 1 to 8 percent of the surface wind velocity. The angle of drift is from 20 to 40 degrees to the right of the wind in the Northern Hemisphere. The speed will be reduced by a factor of four as the ice becomes packed to 90 percent coverage of the ocean. Smooth ice drifts with less speed than rough ice because the wind has greater effect on a rough surface. Rough or hummocked ice usually is thicker and thus has greater inertia; ice of great inertia takes longer to reach the wind factor speed, but after the wind stops the ice continues to move longer than light ice. This phenomenon results in strings of ice floes aligned perpendicular to the wind with small floes packed to windward against larger floes. This wind sorting of ice is a phenomenon of great importance in navigating pack ice, and the strips of open water aligned in a direction approximately perpendicular to the direction of wind explain the successful navigation in ice performed by sailing vessels in the past.

Sediment transport. Both icebergs and pack ice transport sediment in the form of pebbles, cobbles, boulders, and finer material, and even plant and animal life, thousands of miles from their source area. The distribution of icebergs 10,000 years ago, for example, can be inferred from sediments that are widely disseminated on the ocean floor of the North Pacific Ocean as far south as 48° N latitude and in the South Atlantic Ocean as far north as the latitude of the Cape of Good Hope in the Southern Hemisphere. Ice rafting competes with kelp rafting as an explanation for the occurrence of pebbles and boulders on the seafloor as far south as Baja California and other areas of the world ocean.

Bottom freezing of ice shelves and reduction of the ice surface results in migration of sediments and organisms upward. Layering of sediment can be seen in sea-ice floes and icebergs from ice shelves. Fossil penguin bones have been found as far north as 30° S and massive banks of boulders have been noted on the seafloor off South Africa. Thicknesses of as much as 130 centimetres (50 inches) of glacial till have been noted on the ocean floor from discharge by icebergs, apparently during the last 1,000,000 years.

Icebergs are coloured brown, black, and green by a combination of sediment, plankton deposits under the source area ice shelf, and glacial blue ice.

Reliability of radar detection

Iceberg detection and destruction. In the open ocean most ice is seen by radar at ranges depending upon fragment size, but smaller icebergs or growlers can only be detected when the sea surface is calm, and then only at

ranges of about one mile. During slight wind conditions, in particular in heavy seas, the echoes from the waves, known as sea return, may completely mask the echoes of large, potentially very hazardous chunks of ice. In addition, radar return from rain and snow obscures the return from ice in either rain, snow, or fog conditions. Neither radar nor sonar can be relied upon for detection of icebergs and pack ice in choppy seas. The reflectivity of ice and snow to light is great, but reflectivity to radar or short radiowaves is very poor. An iceberg seven metres (22 feet) high cannot be detected with modern equipment if the waves are over four feet high. Various innovations since the inception of radar during World War II have not improved significantly the capabilities of radar to detect icebergs, thus ships are still bound by international agreement to proceed at slow speed in fog.

Sonar is effective in detecting icebergs; however, the range of detection is frequently limited by the water conditions and speed of commercial vessels. The likelihood of insufficient warning for high-speed passenger and cargo ships leaves this mode of detection inadequate. Other suggestions of dropping radio transducers or metal reflectors onto icebergs have the common fault that they are subject to icebergs rolling over and calving, which are frequent occurrences as a dying iceberg melts its way deep into the warmer waters of the shipping lanes.

The problem of iceberg protection, therefore, is one of tracking icebergs as they come down the Labrador Current, and reporting the whereabouts of these floating menaces to all North Atlantic shipping as often as twice daily. This sometimes involves 300 icebergs and requires a team of iceberg experts, oceanographers, aviators, and seamen. During heavy ice conditions two U.S. Coast Guard planes fly six- to eight-hour reconnaissance missions from Argentia, Newfoundland, over the Grand Banks off Newfoundland and contiguous areas. Positions of icebergs are plotted and are correlated with previous positions of the same icebergs or groups of icebergs. When dangerous icebergs approach the shipping lanes, a ship departs for the scene to stand by near the iceberg and warn approaching ships.

Once an iceberg is spotted in a position of threat to ships, it should be of no harm if destroyed. Destroying a 200,000 ton block of ice is, however, a task whose difficulties leave it impractical in most situations. The first successful results on breaking up icebergs by explosives were reported in 1929, when a few hundred pounds of thermite cracked an iceberg into smaller pieces through thermal stress; these then melted more rapidly due to the greater surface area exposed. Similar thermite experiments on two icebergs in 1959 did not corroborate the original findings, however. Attempts at bombing, torpedoing, shelling, and even ramming have been unsuccessful. With twenty 450-kilogram (1,000-pound) bombs acting as direct hits, or as depth charges, it is possible to chip away about 20 percent of a 250,000-ton iceberg. The International Ice Patrol has even tried painting half of an iceberg with lampblack or charcoal to induce thermal stress by heat absorption, but, like other experiments, these attempts were inconclusive.

PACK ICE

Formation and characteristics. On superficial examination, frozen ocean or salt water appears similar to freshwater ice; there are two principal differences, however. First, because the maximum density of seawater occurs below the freezing point, even after freezing the water below the ice will continue to turn over or circulate. As the surface water near the ice becomes colder, it becomes heavier and sinks, resulting in a continuous turnover or vertical circulation of the water beneath the ice. This is a different situation from that which occurs in lakes where, because the maximum density of fresh water is above the freezing point, once ice forms the colder water is lighter than the deeper, somewhat warmer water, and mixing does not occur (see below). A second characteristic is the fact that as seawater freezes, minute pools of salty water called brine pockets are entrapped. The final form and the macroscopic physical properties of the ice are very much dependent upon the concentration of the brine pockets within the ice block. Once an ice field has formed, the physical and chemical properties are not locked in a frozen coffin, but vary as brine pockets migrate through the ice block in response to gravity and thermal gradients.

In the Northern Hemisphere during September and October, the air temperature lowers sufficiently to form a thin sheet of ice. Freezing temperature for average northern ocean salt water of about 3.5 percent salt composition by weight (usually designated 35 parts per thousand) is −1.8° C (28.7° F). The first signs of freezing are changes in the colour and texture of the sea surface as thin, gray-coloured needles and crystal plates form a surface thin sludge. If quiet sea conditions prevail, sheets of crystalline aggregates plate the ocean surface. Initially the ice film is entirely fresh, but as more ice crystals form, pockets of salt water (brine pockets) become entrapped between lamellae (very fine layers) of tiny ice plates. The amount of brine entrapped depends on the temperature of formation and the age of the ice. The shape of the initial ice crystals varies from square discoids to hexagonal dendritic forms. The average width is one inch and the thickness up to about 0.15 centimetres (0.06 inches). During the surface veneer formation stage each grain is free to grow both laterally and vertically until the sheet consolidates. As the ice sheet thickens, the orientation of these grains will change. Due to slight breezes and water motion, the thin sheets of ice jostle about and, after but a few hours, form a field of ice paddies.

The freezing process

The appearance is very much similar to a lily pond completely covered with large gray lily leaves with slightly raised white fringes around their periphery. These disks of ice are known as pancake ice. If the temperature remains below freezing, the pancake ice coalesces as more ice forms, and within a few days the ice cover can be three to four inches thick with a slightly corrugated surface, unless snow prevails, in which case the entire sea area appears as a smooth white plain. As seawater continues to freeze at the bottom edge and sides of ice floes and fields, snow cover increases and the pressures associated with the stresses and strains caused by water and wind movement result in a hummocking and ridge development in some places and open water in other places.

The rate at which the ice forms and thickens depends on the air temperature, ocean turbulent heat flux (mixing conditions), and amount of snow acting as a heat or cold insulator. An empirical formula has been developed and used extensively by the U.S.S.R. and the U.S. to predict ice appearance and growth rate. The equations are based on the simple concept that ice growth is directly related to the length of time during which the air temperature is below the freezing point for seawater. By adding the number of Fahrenheit degrees below the freezing point for each day, a measure of the severity of cold and time of exposure is obtained. This measure is known as the freezing degree days. In north polar regions there are about 8,000 freezing degree days, which is equivalent to four months of −18° C (0° F) air temperature or a mean annual Arctic air temperature of −12° C (10° F; 22° F below freezing).

During ice growth there is surface evaporation and sublimation and bottom ablation when upward heat conduction in the ice is less than ocean upward heat conduction. The balance between ablation and freezing or accumulation results in an equilibrium thickness of about 3.5 metres (11.6 feet) of ice in the Arctic and probably about the same in the Antarctic. The lifetime of this North Polar sea ice is five to eight years, and the lifetime of Antarctic polar class ice found only in the Bellingshausen and Weddell seas is about three years. These lifetime values are related to the rate with which the whole Arctic or Antarctic pack in certain areas moves toward the Equator.

Sometimes early in the season there is sufficient warming and wind induced surface motion to completely disintegrate the ice field. Oftentimes after four or five inches of ice and snow have been formed in a more-or-less uniform manner, a large crack develops, which, through wind and stress, opens into a wide canal commonly known as a lead. This phenomenon is seen frequently in older ice and during the spring breakup. Leads are frequently followed by ships navigating in ice fields, but unfortunately the leads sometimes have a dead end with a very large iceberg blocking the way. With alternating freezing, partial melt-

ing, snow, and wind and swell, the ice field develops over a matter of a few weeks to a month into a six-inch- to two-foot (15- to 61-centimetre) deep ocean cover. At this point the ice field is still navigable by most large vessels; however, if a vessel finds itself in the far north two weeks after the commencement of active surface freezing (*e.g.*, in late October), it is in peril of being locked in for the remainder of the winter.

Saltiness of sea ice

The salt content of seawater as it freezes is always less than that of normal seawater. The amount of salt in the seawater during its first moment in the solid state is dependent upon the rapidity with which the seawater freezes, but in general the salt content is about one-tenth that of seawater. The slower the freezing process, the less the salt content. The most rapid freezing is that which occurs during the first day, and this ice is saltier than underlying ice which forms at the ice–water interface. The salt that remains in the ice is located in tiny pockets of fluid surrounded by normal crystals. These pockets of fluid migrate, mainly by gravity, through the matrix of ice crystals with the result that after a few weeks to a few months, the surface of the ice becomes lower in salt content than the deeper layers. It had once been thought that the difference in temperature, or thermal gradient, was the principal driving force for brine pocket migration; careful experiments indicate, however, that although the thermal gradients are a factor, it is principally gravity that accounts for the movement of these brine pockets. This migration continues throughout the winter.

In the summer, when the ice temperature rises, there is a rapid increase in the migration of salt out of the ice. The sea ice at the surface loses so much salt that it becomes potable and, in fact, is used by Eskimos as a source of fresh water. Salinity reaches a value of less than 0.01 percent. In summary, when first formed the surface layer may have a salinity of 2 to 4 percent, but by April the salinity has dropped to between 0.4 and 0.7 percent, while sea ice that has been through at least one melt season has a salt content below 0.1 percent.

Arctic pack ice. The pack ice of the Northern Hemisphere covers an average area of 10,620,000 square kilometres (4,100,000 square miles; see Figure 18), filling the Arctic Ocean Basin and adjacent North Atlantic Ocean. The polar ice field consists of 4,700,000 square kilometres (1,800,000 square miles), of three- to six-metre- (10- to 20-foot-) thick polar ice that never melts. Infrared imagery from aircraft, however, shows that 10 percent of the polar pack is open water even during winter. Along with Arctic Basin seasonal sea ice, this Arctic pack exudes into the northern Atlantic through two ice streams. The major exit of drifting pack ice from the Arctic Basin is along the eastern side of Greenland, mostly west of Spitsbergen. This ice tongue stretches 2,400 kilometres (1,500 miles) out of the Arctic Ocean and empties a stream of sea ice at a drift rate of eight miles per day. The second icy arm of the north consists of a discharge through the Arctic-Canadian Archipelago and along the eastern American shore. This outpouring of ice is the principal deterrent to easy Northwest Passage ship transit and northern American migration and exploration. During winter, fast ice and

Ice streams into the North Atlantic

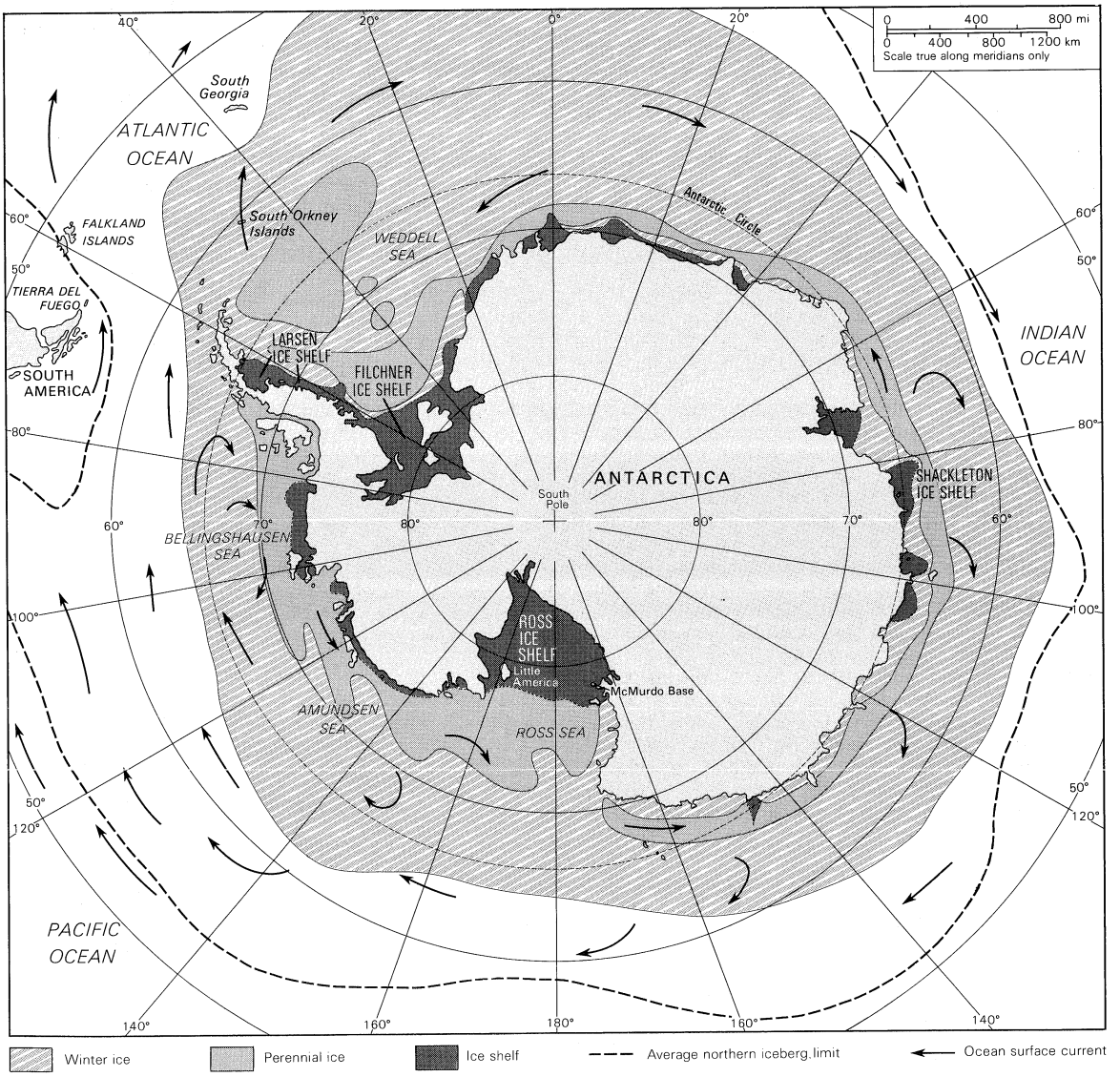


Figure 19: Limits of sea ice and icebergs in the Southern Hemisphere.

local sea ice form along the Siberian coast, Barents and Kara seas, East Greenland, and Labrador coasts down to Newfoundland. The maximum extent of drifting sea ice is about 42° N latitude (about the same latitude as Boston); however, this represents the limit of floating ice pieces and not the hazardous ice-pack edge, which seldom reaches south of Newfoundland. During the summer, the 240-kilometre (150-mile) belt of ice lying along the Labrador coast from Newfoundland northward melts to leave the approaches into Hudson Bay and the Canadian Northwest Territories clear. The motion of the polar ice follows a giant clockwise eddy with a centre 85° N 170° W.

In the North Pacific Ocean comparatively little pack ice and icebergs are encountered. The Bering Sea is clear of pack ice during the northern summer, but commencing in September pack ice forms in bays and is carried through the Bering Strait. In winter and spring pack ice is found as far south as 40° N. This is drifting ice from northern latitudes and the Sea of Okhotsk. During winter, pack ice forms in the northern part of the Sea of Japan.

Antarctic pack ice. Approximately twice as much pack ice forms in the oceans surrounding Antarctica as is found in the Arctic. There are, however, only limited regions in the Bellingshausen and Weddell seas where true polar perennial ice similar to the polar ice cap of the Arctic occurs. The maximum area of Antarctic pack ice is 20,000,000 square kilometres (7,700,000 square miles), or about 8 percent of the Southern Hemisphere.

Figure 19 indicates the extent of Antarctic pack ice. It forms a fairly constant band of drifting sea ice around the continent, with the farthest northern extent occurring at the end of the austral (southern) winter in October. The greatest extension of pack ice in the South Pacific sector is found in about latitude 62° S, and in the South Atlantic pack ice extends to 52° S. The average northern boundary for icebergs is 56° S in the Pacific sector, and 42° S in the South Atlantic. The minimum ice coverage occurs in March, when most of the Antarctic coast is free of ice, with the exception of the Weddell and Bellingshausen seas. The eastern and western coasts of the Weddell Sea are ice free, but the Weddell itself is covered by a slowly (clockwise) revolving pack that seems to have a two-year cycle. The west coast of the Ross Sea is the most predictably open area during the Antarctic summer, and it is here at the approaches to McMurdo Base that most of the Antarctic expeditions have worked their way to the continent. Much of the sea ice encountered in November and December on approaching the continent hundreds of miles from landfall represents ice that formed near the coast from the previous austral winter. In January and February there usually is clear water adjacent to the coast in all sectors around the continent, but this navigable water might be blocked by hundreds of miles of pack ice barrier farther out to sea. The Antarctic freeze-up commences with pack ice formation in the southerly parts of the Weddell Sea followed by pack ice appearance in the Bellingshausen and Ross seas. Beginning in March, ice is formed in sheltered bays, and it extends northward as the sea surface temperature drops. From late February to August, snowfall adds more to the ice thickness than is the case in the Arctic. This is an important difference between Arctic and Antarctic navigation, in that the presence of snow has a cushioning effect and ice breaking is more difficult. By the October maximum the action of wind, sea, and some melting results in very active ice movement. It was in October that Sir Ernest Shackleton's ship "Endurance" was crushed and sank in the Weddell Sea in 1915.

The movement of the ice sheet around the Antarctic continent is from east to west except in the most northern part of the Weddell Sea, where there is a west to east movement making up the northern arm of the Weddell Sea Gyral. This clockwise Weddell eddy has been well documented by the drifts of entrapped ships. The ice drift in the Bellingshausen Sea is less definite. The ship "Antarctic" followed a meandering, aimless course when locked in the ice throughout the 1898–99 winter. From the Ross Sea, ice definitely drifts toward the Weddell Sea under the influence of the prevailing easterly winds near the Antarctic coast.

Sea ice forecasting and reconnaissance. Sea ice reconnaissance and forecasts in the Arctic and Antarctic are conducted by a number of nations, but the principal work is done by the U.S. Navy, either through the Fleet Weather Office or the Ice Forecasting Central of the U.S. Naval Oceanographic Office. The Ice Forecasting Central has for many years supported Arctic and Antarctic research and supply missions through ice reconnaissance and careful short- and long-range forecasting. Additional supportive reconnaissance and regional forecasts are provided by the Canadian government and U.S. Coast Guard ships. Support is even received from commercial airplanes, which frequently spot lonely icebergs far from the area of usual occurrence.

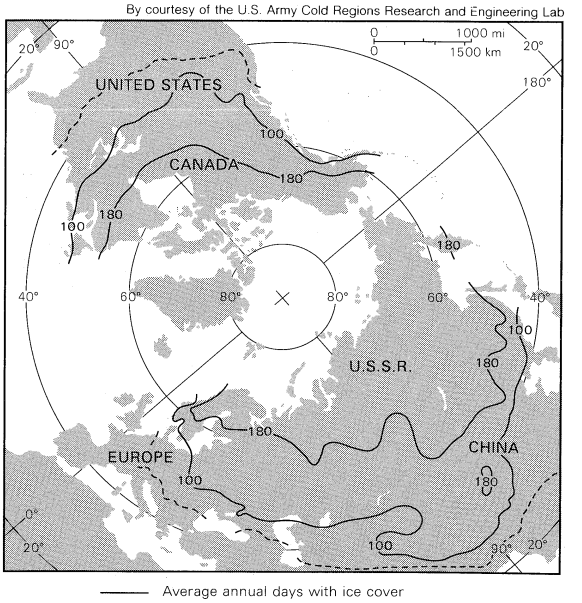
Satellite photographs lack the resolution for day-to-day iceberg reconnaissance; these photographs do show the edge of the sea ice in both polar regions, however. Ice navigation in the Antarctic is similar to navigation for logistic support of military and scientific expeditions in the Arctic. The emphasis is on pack ice distribution and concentration rather than icebergs, which offer little problem to ships once in the ice pack.

(T.F.B./Ed.)

Ice in lakes and rivers

Over much of the world, ice forms on lakes and rivers every year and remains on these surface-water bodies for varying lengths of time. About 80 percent of the conterminous United States experiences freezing or near-freezing water temperatures in an average year; freezing surface-water temperatures persist for more than 100 days in an average year over about 25 percent of the same area (see Figure 20). The surface waters of 48 percent of the Northern Hemisphere (49,000,000 square kilometres or, 19,000,000 square miles) freeze annually. Data are less complete for the Southern Hemisphere, but with the exception of southern South America and Antarctica, the landmasses there do not extend above 40° S latitude, and surface-water freezing is uncommon except at high altitudes. Lakes in Antarctica are completely ice covered perennially; and in Greenland, Canada, Siberia, and other high-Arctic lands, lakes may have a partial ice cover throughout the year. In some areas where freezing is general, certain lakes and river reaches may avoid freezing because of thermal springs, significant influxes of above-freezing groundwater (which is usually at a temperature close to the mean an-

By courtesy of the U.S. Army Cold Regions Research and Engineering Lab



Average annual days with ice cover

Approximate southern boundary of the occurrence of annual lake and river freezing

Figure 20: Distribution of ice cover in the Northern Hemisphere.

nual temperature of a region), unusually high salt content, or inputs of water artificially heated in industrial cooling processes. Extreme turbulence also may prevent ice-cover formation, particularly in certain reaches of rivers or at inlets or outlets of lakes. Very deep lakes may not freeze because winter cooling does not sufficiently lower the temperatures of their large water volumes.

Economic and practical importance

Lake and river ice is of great practical and economic importance. Until the development of mechanical refrigeration, ice was annually cut from lakes and rivers and used to refrigerate food; the practice persists in some regions. The greatest economic impact of freshwater ice is doubtless in preventing navigation on inland waterways, however. In North America the annual loss attributed to inability to navigate frozen waterways approaches $200,000,000 on the Mississippi River and the Great Lakes–St. Lawrence Seaway combined. In Siberia, where other forms of transportation are less well developed, the losses due to the short shipping season are of even greater relative importance. Losses of life and severe economic losses are associated with flooding due to ice jams and river icings, which are annual occurrences in many mid- and high-latitude regions.

River ice also greatly reduces the exchange of oxygen between the air and the water, markedly slowing the natural re-oxidation of water that is transporting biochemically degradable wastes. Freezing also reduces streamflow quantities available for water supply, waste dilution, and power generation; the ice cover represents water in storage and, hence, water that is temporarily unavailable. On a 30-metre- (100-foot-) wide river with an ice cover 30 centimetres (one foot) thick, about 9,000 cubic metres (318,000 cubic feet, or 7.3 acre-feet) of water is in storage for each kilometre (0.6 mile) of river length.

The thermal expansion and contraction of river and lake ice generates significant stresses against dams, piers, and other structures, which must be accounted for in engineering designs. During breakup ice jams are often major causes of damaging floods; floating ice cakes associated with high river flows have been known to dislodge bridge superstructures from their piers. Hydroelectric power plants may cease generation because of the accumulation of ice on intake trash racks or screens, completely blocking river flow through the turbines. Frozen lakes and rivers provide some positive economic benefits: recreational resources for skaters, ice fishermen, and iceboaters. In remote areas with significant freezing, lake ice generally provides the most suitable landing sites for aircraft.

Effects on flora and fauna

Lake and river flora and fauna are affected by freezing in several ways. Surface feeding by fish is prevented, and waterfowl can no longer rest and feed when an ice cover forms. Light penetration is restricted and photosynthesis virtually ceases. Shore vegetation may be injured or destroyed by ice formation and subsequent vertical and horizontal movement of the ice cover. Several studies have shown that ice formation on river bottoms dislodges and disperses bottom organisms; however, the ecological effects of this are apparently not great.

ICE IN LAKES

Ice formation. The temperature of most lakes falls largely because of heat loss from the lake surface; it rises when the surface gains heat. This heat exchange occurs by absorption and reflection of shortwave (solar) radiation; absorption, reflection, and emission of long-wave (thermal) radiation; evaporation; and convection of air. Precipitation, inflow of surface and ground waters, and geothermal heat flow generally account for only minor heat exchanges. Thus, for a given water temperature, the rates of radiation exchange, evaporation, and conduction-convection are controlled by weather conditions. When the difference between air and water temperature is less than 5° C (9° F) at high wind speeds and less than 20° C (36° F) at low wind speeds, long-wave radiation and evaporation principally account for the heat loss. With larger temperature differences, conductive-convective heat exchange dominates.

As a lake cools from a temperature exceeding 4° C (39° F), the water at the surface loses heat, becomes denser than the water below it, and sinks. This process continues until all water in the lake is at 4° C, which is the temperature of maximum density for water. Wind not only accelerates surface heat loss but promotes cooling by increasing internal circulation. Further cooling at the surface produces a less dense layer, which tends to remain at the surface. Cooling of the surface layer proceeds until a temperature a few tenths of a degree below the freezing point, 0° C (32° F), is reached.

Supercooling and ice-crystal formation

Supercooling is required so that the latent heat of freezing can be carried away. In water at exactly 0° C, any freezing will liberate heat (80 calories per gram, or 144 British Thermal Units [BTU] per pound, of water freezing), thus warming the unfrozen water above 0° C and inhibiting further freezing. The degree of supercooling required depends on the character of nuclei present. Nuclei generally are existing ice crystals or dissolved or suspended minerals that have crystal structures like that of ice; such impurities are virtually always present in natural water. If nuclei are not present, small ice crystals eventually form, but at greater degrees of supercooling.

In a small isolated body of still water, spontaneously forming ice crystals initially occur in the form of disks a few millimetres in diameter. The disk shape is dictated by the condition of minimum surface energy and by the crystallographic structure of ice. As the disk enlarges, the curvature of the circumference decreases, and the heat of freezing cannot be efficiently disposed of. At this point, minor irregularities in the disk begin to enlarge and grow into a system of branching needles, called dendrites.

Crystal growth is fastest in the direction in which the heat of freezing is most readily dissipated. For a floating crystal, this direction is parallel to the surface (Figure 21). On still water, the surface can be completely covered with

From Ven Te Chow (ed.), *Handbook of Applied Hydrology*, Copyright 1964; used with permission of McGraw Hill Book Company

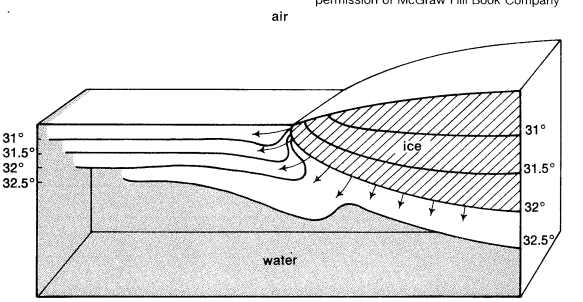


Figure 21: Growth of ice crystal in supercooled water. Arrows indicate direction and amount of ice-crystal growth.

a thin skim of ice, called sheet ice, very quickly. Where there is some agitation of the surface by wind, a thin skim may be broken by wave action. Individual pieces are then subject to abrasion before finally fusing together to form one type of agglomeratic ice. A second type of agglomeratic ice forms when snow falls on a water surface that is very near the freezing point. The resulting irregular masses of slush will subsequently freeze together if cooling continues. On large water bodies, agglomeratic ice generally forms the first complete ice cover. On most lakes, a skim of ice, or thin agglomeratic ice, forms and dissipates up to several times before the "permanent" seasonal ice cover is established.

Ice growth. After a complete ice cover is established, it thickens largely by freezing on the underside. This growth occurs in the form of downward-growing columns. Dissolved impurities and air bubbles, which are largely rejected from the ice during the freezing process, tend to be trapped and concentrated along the boundaries of the columns. The faster this downward freezing occurs, the greater the incorporation of impurities along these crystal boundaries. Additional thickening can occur when an ice cover cracks from the weight of snow, allowing water to flow over the surface. When this mixture freezes, a condition called snow ice is produced.

Equilibrium thickness of ice

The rate of accretion of ice on the underside is controlled by the rate of heat flow, and, hence, the temperature gradient through the ice (and perhaps snow on the surface).

The equilibrium thickness of an ice sheet, in the absence of snow cover, depends on the air temperature and time as shown by the following formula:

$$x = \sqrt{\frac{2k_i(T_o - T_a)t}{Q_L \rho_i}},$$

in which x is the equilibrium thickness, k_i is the thermal conductivity of ice (.00535 cal/cm · sec · ° C, or .000358 BTU/ft · sec · ° F), T_0 is the freezing temperature of ice, T_a is the air temperature, t is time, Q_L is the latent heat of freezing, or heat emitted by crystallization of a unit amount (80 calories per gram or 144 BTU per pound), and ρ_i is the density of ice (0.92 grams per cubic centimetre or 57.4 pounds per cubic foot).

The effect of snow on an ice layer is to increase the time required for the ice to reach a given thickness, because snow is an efficient insulator. This effect is taken into account in an empirical equation for ice-thickness prediction:

$$x = \beta(1.06\sqrt{D}),$$

in which x is ice thickness in inches; β is a coefficient ranging from 0.65 to 1.00, depending upon the effects of snow cover, winds, and other environmental factors; and D is accumulated degree-days below 0° C since freeze-up. Hence, the ice thickness varies with time and temperature but also with snow cover and other factors whose effects are assigned numerical values between 0.65 and 1.00.

Several studies have indicated that there is a gradual increase in water temperature at depth after ice-cover formation. This has been attributed to five possible causes: (1) solar radiation transmitted through the ice cover and absorbed by the water; (2) transfer of heat stored in lake-bottom sediments to bottom water; (3) inflow of warm groundwater or surface water; (4) oxidation of organic material in bottom sediments; and (5) geothermal heat flow. The few studies to date suggest that release of heat stored in bottom sediments is the principal cause for at-depth temperature increase.

Table 5: Time Required for Ice-Sheet Formation

thickness (cm)	air temperature			
	−1°C	−5°C	−10°C	−20°C
1	1.9 hours	0.38 hour	0.19 hour	0.10 hour
2	7.6 hours	1.5 hours	0.76 hour	0.38 hour
5	2.0 days	9.5 hours	4.8 hours	2.4 hours
10	7.9 days	1.6 days	19.0 hours	9.5 hours
20	32 days	6.3 days	3.2 days	1.6 days
50	198 days	40 days	20 days	9.9 days

Ice decay. The presence of water at above-freezing temperatures beneath the ice indicates an upward flow of heat. The net heat flow to the ice–water interface, however, remains negative (*i.e.*, heat is removed and freezing occurs as indicated above) until air temperatures rise to approximately the freezing point. At such time melting occurs at both upper and lower surfaces, but is much more intense on the upper surface. Solar radiation, conduction from the air, and heat derived from rain are the principal energy sources leading to decay and dissipation of the ice cover. Melting is most rapid along crystal boundaries, due to the concentration of impurities (which lowers the freezing point of a solution) and to the presence of imperfections in the crystal lattice that are concentrated at such boundaries. If the crystals are columnar, so-called candle ice forms as a result. Thus, an ice cover may weaken significantly during the early melt season without a corresponding decrease in thickness. As surface melting proceeds, the presence of water on and in the snow or ice cover rapidly reduces the albedo of the surface. This increases the rate at which solar radiation is absorbed (from perhaps 10 percent absorbed for dry snow to 70 percent for slush) and thus hastens melting.

Factors influencing surficial melting

Melting generally accelerates along the lakeshore and near an inlet or outlet stream. Near-shore melting is a response to conduction and long-wave radiation from shoreline materials and vegetation, which absorb more solar radiation than does ice or snow. An ice cover may thaw completely around the shore and become completely freed; wind and wave action at the edges may then contribute to dissipation. Flowing water at inlets or outlets is a constant source of heat transport to the underside of the ice cover, and ice is generally thinner in such places.

It has been observed that water temperatures begin to fall as melting progresses, particularly if there are open zones along the lakeshore. Any of three processes (or a combination of them) may account for this: (1) near-freezing water released at the water–ice interface as the ice melts; (2) inflow of meltwater down through cracks in the ice cover; and (3) inflow of meltwater from streams and lakeshore. The relative importance of these processes probably varies from lake to lake, but few data are available on which to base quantitative statements. Once the ice cover is gone, water temperatures again begin to increase by solar heating.

Forecasts of ice breakup are based on air-temperature indices established for particular locations.

Geographic distribution. In Canada, the average date of the first appearance of ice in lakes (freeze-up) ranges from September 1 in the extreme north to December 20 in the Great Lakes region. Complete freeze-over of lakes generally occurs about ten days after first ice. Although freeze-over and freeze-up times are generally correlative with latitude, both are strongly modified by altitude and distance from the ocean. In general, areas with continental climates experience freeze-over and freeze-up considerably earlier than do areas with maritime-influenced climates at the same latitude. At latitude 60° N, for example, average freeze-over dates range from January 1 near the Pacific coast to November 1 in the area around Hudson Bay. Because the moderating influence of the ocean is less pronounced along Canada's east coast (because of the cold Labrador Current), average freeze-over date is approximately November 15 at the same latitude there.

Dates of freeze-up

In the United States east of the Rockies, freeze-up dates range from November 15 in the north central states to about January 1 at latitude 40° N and have a fairly consistent relationship to latitude. West of the Rockies, however, patterns are more complicated because of topography and the influence of strong maritime climates in the Pacific Northwest, and data are insufficient to reveal consistent patterns. In much of the area, freezing does not occur every year.

Mean dates of lake freezing in the Soviet Union range from about October 1 at or above latitude 72° N to mid-January at latitude 45° N. East of the Urals, freeze-over dates are generally correlated with latitude, with minor modifications in areas of strong continentality—*i.e.*, where climate is governed principally by the presence of large land areas. In the western portion of the U.S.S.R., dates of freeze-over become later as one progresses westward; at 60° N, they range from November 1 at the Urals to mid-December at the Baltic Sea.

Data on the relation between lake-freezing dates and altitude are scarce throughout the world. In general, temperatures decrease about 6° C for each 1,000 metres of increase in elevation (about 3.5° F per 1,000 feet), but quite significant variations should be expected, depending on local climatic factors (the microclimate).

Dates of breakup

In Canada, the first distinct signs of breakup on lakes (surface melting or open water) appear on the average as early as March 10 in the extreme southern Great Lakes area to mid-July in the Arctic archipelago. Breakup may occur from 10 days to a month after the day on which average air temperature first rises to 0° C (32° F). Dates on which ice completely disappears from lakes occur usually from five to 30 days after the initial signs of breakup: April 1 in the far south to August 1 in the extreme north. As with freezing dates, the pattern of average breakup and clearing dates is strongly modified by continentality. The total ice-free season on Canadian lakes lasts for as few as 30 days in the extreme northern islands to as many as 270 days in the Great Lakes.

Breakup in lakes in the United States begins in early March near the southern limit of annual freezing (about

40° N, east of the Rockies) and in late April in northern Minnesota and Maine. Again, because freezing is not an annual occurrence in much of the area west of the Rockies and because irregular patterns result from topographic effects, no generalizations can be made.

In the Soviet Union, average breakup dates extend from late June in the Arctic Ocean islands to mid-February near latitude 42° N.

ICE IN RIVERS

Formation and growth. Temperature changes in rivers, as in lakes, are caused principally by heat exchange across the surface, and the same cooling processes are important. The important difference between freezing processes in rivers and lakes is that, unlike lakes, virtually all natural streams are turbulent. This means that the flow is well mixed—*i.e.,* that any parcel of water is continually changing its vertical, as well as its horizontal, position in the flow. Because of that turbulence, temperature stratifications cannot arise, and sheet ice cannot form in streams, except along the edges where velocities are at or close to zero. Even in deep rivers, water temperatures are usually at or very close to the freezing point throughout the vertical profile from surface to river bottom during the ice season.

In rivers, only small degrees of supercooling are needed to form ice, as turbulence quickly carries away the heat of freezing, losing it to the atmosphere at the river surface.

Frazil ice and anchor ice

If cooling is rapid and turbulent water is supercooled a few hundredths to a tenth of a degree Celsius, small disks of ice, called frazil ice, begin to form. If ice is already present as sheet ice near the edges of the flow, these disks form at the water–ice interface at temperatures of −0.01° to −0.03° C (31.98° to 31.95° F) and quickly break off and become suspended in the water. If no ice is present, the frazil disks form throughout the flow, apparently at greater degrees of supercooling (−0.04° to −0.08° C [31.93° to 31.86° F]).

Although the frazil ice has the same density as normal ice (0.917 grams per cubic centimetre), turbulence is commonly sufficient to keep it in suspension, so that it does not collect at the surface. The frazil disks have a high degree of adhesiveness for each other and for certain surfaces. Consequently, they form large, slushy, irregular masses that may subsequently adhere to the bottom as anchor ice. As frazil ice collects on a rock on a stream bottom, the buoyancy of the adhering ice may be sufficient to lift the rock and carry it downstream. That adhesive property of frazil ice causes trouble at hydroelectric plants; frazil ice collects on the trash racks of intakes to the turbines, where it may quickly accumulate to block the flow and cause the plant to shut down. The sequence of events leading to establishment of a complete ice cover on a river varies from area to area and from year to year, depending principally on channel configuration and hydrologic and meteorologic conditions. Generally, the process begins with the formation of sheet ice or agglomeratic ice or both where velocities are low—*e.g.,* along the banks and in backwaters. Sheet ice may continue to form over the entire river width if velocities are everywhere less than about 0.5 metre (1.6 feet) per second.

In faster flowing reaches, where turbulence prevents sheet-ice formation, frazil ice forms instead. As the frazil-ice particles are carried downstream and encounter stretches of lesser turbulence, they may float to the surface and collect in slushy masses 30 centimetres (one foot) to a metre (3.2 feet) or more across, forming shell ice (Figure 22). Along with pieces of sheet ice that have been broken from the shore by buoyancy forces as velocities change and river levels vary, such slushy masses generally collect at a channel constriction or where border sheet ice extends across the surface, forming an ice bridge.

Groundwater and surface water discharges

If upstream of the ice bridge the river's flow slows to less than a critical value—depending on the thickness of the bridge and the flow depth—the ice cover will progress upstream. This critical velocity, V_c, is given by the formula

$$V_c = \left(1 - \frac{h}{H}\right)\sqrt{2g\left(\frac{\rho - \rho_i}{\rho}\right)h}$$



Figure 22: Shell ice flow on the Athabasca River, Alberta.
Alberta Government Photograph

in which g is acceleration of gravity, ρ and ρ_i are the densities of water and ice, respectively, h is thickness of the ice bridge, and H is flow depth just upstream of the ice bridge. If flow velocity exceeds V_c, the ice floes arriving at the upstream edge of the bridge will be carried under it, along with frazil disks and needles, if present. Some of this ice may adhere to the underside of the ice bridge, causing it to thicken. This thickening increases the value of h in the above equation and also reduces the flow velocity upstream, so that eventually the critical velocity for upstream growth is reached and the ice cover can progress. The rate of thickening depends upon the amount of ice carried by the flow per unit time, the flow velocity under the ice, and the roughness of the underside of the ice bridge. Thickening of an existing ice bridge can also occur by plastic deformation or by breaking and refreezing if stresses, due mainly to hydraulic forces at the edge of the ice, are sufficient.

As an ice bridge extends upstream by accumulation of floes, turbulence is reduced in the surface layers of water among the floes and freezing proceeds by cooling at the surface, so that the ice cover becomes complete. Further, river discharge is generally decreased as more water goes into storage as ice, or is stored in backwaters caused by local ice formation. This reduces flow velocities and thereby makes conditions for upstream ice growth more favourable.

In some cases, shell-ice and sheet-ice floes may accumulate at a natural constriction or at such an artificial obstacle as bridge piers to such an extent that an ice dam may form. Such a dam may extend all the way to the river bottom and may cause backwaters extending great distances upstream, perhaps giving rise to flooding.

On streams large enough to provide a depth of flow greater than the ice thickness, a complete ice cover is suspended from the channel banks and floats on the river surface. The ice generally is flexible enough to rise and fall with changes in river level. If a change in level is rapid, or if the weight of snow accumulating on the ice is sufficient, cracks may develop, allowing water to spill over the surface where it subsequently freezes and adds to the thickness of the ice cover.

Even in the absence of cracking, the ice thickness changes during the course of a winter in response to weather and flow conditions. Thinning occurs if water discharge increases, if atmospheric and radiational processes supply heat to the ice sufficient to cause surface melting and evaporation, or if sufficient snow is deposited. The snow acts as an insulator and decreases the flow of heat from the water through the ice–snow layer to the air; consequently, melting occurs at the interface of ice and water.

The underside of a river-ice cover is generally rough, especially where ice dams have formed. Even where the ice surface is relatively smooth, however, observations indicate that streamlined forms closely resembling the ripples and dunes of alluvial streambeds are present, at least where velocities are greater than 0.5 metre (1.6 feet) per second.

Contrary to earlier speculations, the roughness of these forms shows a tendency to increase through the winter.

Groundwater and surface water discharges

Many arctic and subarctic streams experience freezing that arises differently from the more general sequence of events in temperate zone areas. In those extremely cold regions, winter freezing of the ground is profound, and groundwater flow is reduced to insignificant amounts. This is particularly true where permafrost is present and freezing from the surface eventually reaches to or nearly to the permafrost table (see above). Groundwater flow, which maintains streamflow during periods of snow cover, therefore, is impeded; and stream discharges dwindle or halt altogether. Under these conditions, and especially in the typically wide, shallow rivers of glacial areas, rivers may freeze solid early in the winter. When drainage is thus blocked, pressures in the flowing groundwater or surface water may increase to such an extent that the water breaks through to the surface near the upstream end of a solidly frozen reach. This water spreads out over the surface and freezes. Successive sheets of overflow ice build up icings (German *aufeis;* Russian *naled*) as much as three metres (10 feet) thick and of great extent. Such river icings may persist well into the summer, and even through several years in high arctic areas, and may block roads at bridges or culverts. They can also give rise to ice jams and flooding during breakup.

Decay and ice jams. Breakup of river ice occurs as a result of increased solar radiation, which melts and weakens the ice cover as it does in lakes, and of the increased discharge accompanying snowmelt and spring rains. As river levels rise and flow velocities increase, the weakened ice is broken loose and begins to move downstream. As the ice cover is thus diminished, an additional increase in flow occurs when the water held in storage by the retarding effect of the ice cover is released. In most cases, river breakup is a sudden and dramatic event, and only a few hours may be required to clear large reaches.

Ice jams, or gorges, of greater or lesser magnitude almost invariably accompany breakup. They occur at natural or artificial channel constrictions, as during ice-cover formation, or at an ice bridge or icing where breakup has not occurred. Ice jams during breakup usually cause much more serious problems than during freeze-up, as much greater amounts and thicknesses of ice are present and river discharges are typically near their annual peaks.

Water surges and ice floe damage

When the collected ice floes form dams, the flowing water may spread widely over the upstream flood plain. The damming action will continue until sufficient pressure builds up to free the jam or until the water finds another channel and bypasses the obstruction. Either result generally releases a sudden surge of water and ice downstream of the jam, possibly doing serious damage to bridges and flood plain structures and causing severe bank and channel erosion. The water and ice released at one jam may accumulate at another downstream or may be of sufficient force to dislodge a downstream jam, causing a chain reaction that results in serious flooding.

The ice floes themselves can do tremendous damage in eroding banks and in dislodging large structures—a three-metre- (10-foot-) square ice floe 30 centimetres (one foot) thick weighs about 2,700 kilograms (5,900 pounds). High, ice-laden flows have been known to remove bridge superstructures completely from their piers.

River stretches that have low slopes, broad channels, and numerous bends, islands, and constrictions are most susceptible to ice-jam formation. Large rivers with northerly courses are subject to jams because breakup occurs in the headwater areas to the south before it occurs in the lower reaches. Ice jams commonly form at the same places year after year, but it is very difficult to predict their severity. In a given river reach, the following conditions contribute to severe jamming: (1) low water levels at freeze-up, resulting in a restricted flow section; (2) a cold, prolonged winter, resulting in thick ice; and (3) a rapid thaw, resulting in the production of large amounts of snowmelt runoff in a short period of time.

Forecasting spring ice jams and related flooding is an inexact science. Prediction schemes have been based on such data as the severity of ice-jam formation during the preceding freeze-up season; observations of river level in the early stages of breakup; and observations and forecasts of air temperature, ice thickness, amount of water in the snowpack, and the intensity of solar heating. Some degree of success has been reported with these methods, and predictions will undoubtedly improve as observational networks improve and experience accumulates.

Ice modification. Because of the great human and economic losses associated with ice-jam flooding and restrictions on navigation during the ice season, operations have been undertaken to alleviate ice jams and maintain navigation routes, particularly in the northern United States, Canada, northern Europe, and the Soviet Union. The use of icebreaking ships and tugs is widespread in the U.S.S.R., northern Europe, and the Great Lakes–St. Lawrence Seaway, both to delay freeze-up and to accelerate and assure a jam-free breakup.

Dusting ice with carbon black or other material that increases the absorption of solar radiation by the ice surface (or decreases its albedo [reflectivity]) accelerates melting. In this increasingly popular procedure, the material generally is applied as a powder from an airplane. The method is effective only during the normal melt season; if applied too early, the resulting meltwater may refreeze, or the black material may be covered by snow, rendering it ineffective. The application is usually made on river ice and not specifically on ice jams.

Explosives, diversion structures, and thermal pollution

Explosives, such as dynamite, are used extensively in ice-jam removal. An ice-free channel below the jam and a sufficient flow velocity to carry away the loosened ice are required for success in removing an ice jam by explosives. Blasting operations are most effective when begun soon after the jam forms; underwater charges are more effective than surface charges. Aerial bombing and rocketry have also proved of some success in destroying ice jams.

Ice retention and diversion structures, including especially designed dams and ice booms, are also used to prevent ice jams. Basically, these structures are designed to retain ice so that it will melt where it has formed or to collect it so that it will melt where damaging floods will not occur, while allowing the water to flow past. In 1964 a three-kilometre- (two-mile-) long ice boom was installed on Lake Erie between Buffalo, New York, and Fort Erie, Ontario, at the head of the Niagara River. The boom successfully prevented large quantities of ice from moving from the lake to the river, particularly during periods of high westerly winds.

Changes in water level achieved by opening and closing sluice gates of dams or other flow-control structures have been used to break up an ice cover or to float away ice that would otherwise accumulate in a jam. River modification, including channel dredging, widening, and straightening, are sometimes used to eliminate places where jams frequently form.

Steam-electric power plants, many industrial processes, and sewage-treatment plants release water that has been artificially heated. It has been suggested that this warm effluent ("thermal pollution") could be used to keep river reaches free of ice and thereby extend navigation seasons and reduce ice-jam flooding hazards. Although no conscious efforts have been made to manage water temperatures deliberately by proper positioning of thermal pollution sources, changes in river temperature regimes and ice conditions that have resulted from such sources have been widely noted.

A recent study has investigated the feasibility of extending the navigation season of the Great Lakes–St. Lawrence Seaway by judicious location of nuclear electrical generating plants. Economic gains amounting to $100 million annually would accrue if year-round navigation could be maintained there.

Calculations predict that a nuclear reactor of the size typically constructed during the 1970s could maintain between one-half and all of the 32-kilometre- (20-mile-) long South Shore Canal (above Montreal) in an ice-free condition, depending on severity of winters. With reactors two to three times as big, correspondingly longer reaches of the seaway could be kept ice-free. Those figures indicate the definite feasibility of using nuclear-reactor power-

plant cooling effluent to keep at least selected portions of the St. Lawrence Seaway ice-free for much of the winter. The prevention of ice formation in the vicinity of canal locks would be especially beneficial. The ecological effects of such a plan must first be determined, however.

Geographic distribution. The general pattern and dates of river freezing are similar to those for lakes. Complete freeze-over in Canada usually begins in late September in the extreme north, and on the average about December 20 in the Great Lakes region and January 10 in western British Columbia. A general correlation with latitude exists, but the pattern is less regular than for lakes because of the many local factors that affect the freezing of rivers. Over most of Canada, rivers freeze over from zero to 40 days after lakes, but there are areas where rivers freeze as much as 30 days earlier.

In the United States and the Soviet Union, the dates and patterns of river freeze-up and breakup are essentially the same as previously cited for lakes. Again, local variations are to be expected. On the large northward-flowing Soviet rivers—the Ob, the Yenisey, and the Lena—freeze-over occurs one to three weeks later than the dates previously given for a specific latitude.

The initial breakup of ice in Canadian rivers begins on the average as early as March 1 in western British Columbia and about March 20 in the Great Lakes and averages about June 10 north of Hudson Bay. Clearing of ice from rivers occurs 10 to 20 days later. (S.L.D.)

BIBLIOGRAPHY

General works: P.A. SHUMSKIY, *Principles of Structural Glaciology* (1964; originally published in Russian, 1955), a reference work mainly on the petrography of ice, including sections on ice physics and chemistry, ice formation, ice metamorphism, and methods of study; M.F. MEIER, "Ice and Glaciers," in V.T. CHOW (ed.), *Handbook of Applied Hydrology,* sect. 16 (1964), a discussion of ice and its properties as well as ice formations. See also W.D. KINGERY (ed.), *Ice and Snow: Properties, Processes and Applications* (1963); and E.R. POUNDER, *Physics of Ice* (1965).

Glaciers: L. LLIBOUTRY, *Traité de glaciologie,* 2 vol. (1965), an authoritative work on the subject of glaciology, with an especially thorough treatment of glaciers and ice sheets emphasizing the geophysical, quantitative approach; W.S.B. PATERSON, *The Physics of Glaciers,* 2nd ed. (1981), an excellent, clearly written text on glaciers and ice sheets emphasizing the mechanics of flow. C.R. BENTLEY, "The Structure of Antarctica and Its Ice Cover," H.C. HOINKES, "Glacial Meteorology," P.A. SHUMSKIY, A.N. KRENKE, and I.A. ZOTIKOV, "Ice and Its Changes," and M.J. RUBIN, "Antarctic Weather and Climate," in H. ODISHAW (ed.), *Research in Geophysics,* vol. 2, ch. 14–17 (1964), excellent reviews of the state of knowledge of glacier meteorology and the Antarctic Ice Sheet as of the end of the International Geophysical Year, by leading authorities; M.F. MEIER, "Ice and Glaciers," in V.T. CHOW (ed.), *Handbook of Applied Hydrology,* sect. 16 (1964), a discussion of glaciers at an intermediate technical level, with some emphasis on glacier hydrology; R.P. SHARP, *Glaciers* (1960), an excellent, well-illustrated introduction to the subject of glacier study.

Permafrost: R.J.E. BROWN, *Permafrost in Canada* (1970), the only comprehensive book on permafrost in Canada; D.M. HOPKINS *et al.,* "Permafrost and Ground Water in Alaska," *Prof. Pap. U.S. Geol. Surv. 264-F,* pp. 113–146 (1955), a good review of permafrost effect on groundwater and methods of determining permafrost from aerial photographs; A.H. LACHENBRUCH, "Mechanics of Thermal Contraction Cracks and Ice Wedge Polygons in Permafrost," *Spec. Pap. Geol. Soc. Am. 70* (1962), classic paper on quantitative interpretation of the formation of ice-wedge polygons; NATIONAL ACADEMY OF SCIENCES-NATIONAL RESEARCH COUNCIL, *Proc. Intern. Permafrost Conf., 11–15 Nov., 1963,* Pub. No. 1287 (1966), an extensive collection of papers on permafrost presented at the First International Permafrost Conference; T.L. PEWE (ed.), *The Periglacial Environment: Past and Present* (1969), a thoughtful approach to discussion of present and past permafrost throughout the world, and with R.E. CHURCH and M.J. ANDERSON, "Origin and Paleoclimatic Significance of Large Scale Polygons in the Donnelly Dome Area, Alaska," *Spec. Pap. Geol. Soc. Am. 109* (1969), a thorough discussion of "fossil" ice-wedge polygons and relict permafrost in central Alaska with paleoclimatic interpretations; A.I. POPOV, *Podzemnyĭled,* 3 vol. (1965), an extensive review with informative papers dealing with ground ice and permafrost (in Russian, with preface and table of contents also in English); I.A. TYUTYUNOV, *An Introduction to the Theory of the Formation of Frozen Rocks,* trans. from the Russian by J.O.H. MUHLHAUS (1964), a comprehensive summary of the mechanics and chemistry of the freezing process and permafrost formation.

Icebergs and pack ice: The published literature on icebergs is found mainly in U.S. and Soviet journals, and atlases prepared by the U.S. Coast Guard and U.S. Navy for Arctic and Antarctic military and scientific maritime resupply expeditions. A classic discussion of icebergs and sea ice in the Arctic Ocean is by the late ADM. E.H. SMITH, *The Marion Expedition to Davis Strait and Baffin Bay: Scientific Results,* part. 3 (1931); see also FRIDTJOF NANSEN'S treatment of the oceanography of the north polar basin in *The Norwegian North Polar Expedition, 1893–1896: Scientific Results,* vol. 3 (1902, reprinted 1969), and *Farthest North,* 2 vol. (1898). A great deal of Soviet work on both sea ice and icebergs was done under the leadership of N.N. ZUBOV and published in a book that has been translated by the U.S. Office of Naval Research, *Arctic Ice* (1948). The importance of ice in the water budget of the planet may be studied by consulting J.L. DYSON, *The World of Ice* (1962). *Dynamics of Snow and Ice Masses,* ed. by SAMUEL C. COLBECK (1980), is a comprehensive review of the subject. *Sea Ice Processes and Models: Proceedings of the Arctic Ice Dynamicas Joint Experiment and International Commission on Snow and Ice Symposium,* ed. by ROBERT C. PRITCHARD (1980), reports field observations and the development of models in an effort to establish usable flow laws for sea ice. See also U.S. NAVAL OCEANOGRAPHIC OFFICE (HYDROGRAPHIC OFFICE), *A Functional Glossary of Ice Terminology,* H.O. Pub. No. 609 (1952); and WORLD METEOROLOGICAL ORGANIZATION, *Sea-Ice Nomenclature* (1970).

The age of icebergs can be understood from consulting W. DANSGAARD *et al.,* "One Thousand Centuries of Climatic Record from Camp Century on Greenland Ice Sheet," *Science,* 166:377–381 (1969); and P.F. SCHOLANDER *et al.,* "Composition of Gas Bubbles in Greenland Icebergs," *J. Glaciol.,* 3: 813–822 (1961). A major reference work on the proposed use of icebergs as a freshwater source is *Iceberg Utilization: Proceedings of the First International Conference and Workshops on Iceberg Utilization,* ed. by A.A. HUSSEINY (1978).

Ice in lakes and rivers: B. MICHEL, "Winter Regime of Rivers and Lakes," *United States Army Cold Regions Research and Engineering Laboratory Science and Engineering Monograph III-B1a* (1971), provides a comprehensive treatment of freshwater ice; M.F. MEIER "Ice and Glaciers," in V.T. CHOW (ed.), *Handbook of Applied Hydrology,* (1964), gives a brief but accurate treatment of ice structure and physical properties and the growth, formation, and decay of river and lake ice; J.H. ZUMBERGE and J.C. AYERS, "Hydrology of Lakes and Swamps," also in *Handbook of Applied Hydrology,* summarizes adequately the formation, growth, and decay of lake ice and gives some useful references; S.J. BOLSENGA, "River Ice Jams," *United States Lake Survey Research Report 5-5* (1968), is a thorough compilation of all aspects of ice jams, with extensive extracts from the pertinent literature. For information regarding the geographic distribution of lake and river ice, see the following works: U.S. NAVAL OCEANOGRAPHIC OFFICE (HYDROGRAPHIC OFFICE), *Ice Atlas of the Northern Hemisphere* (1946), gives maps showing dates of opening and closing of navigable waters due to ice. The *Fiziko-Geograficheskiy Atlas Mira,* published by the ACADEMY OF SCIENCES OF THE U.S.S.R. (1964), provides similar but more up-to-date information for the U.S.S.R. W.T.R. ALLEN, "Break-Up and Freeze-Up Dates in Canada," *Canada Department of Transport, Meteorological Branch Circular 4116* (1964), summarizes the data for rivers and lakes in the form of maps for that country.

Iceland

Iceland (Icelandic: Ísland), an island country located in the North Atlantic Ocean, is a land of vivid contrasts. Sparkling glaciers lie across its ruggedly beautiful mountain ranges, while a vast quantity of subterranean thermal activity makes Iceland one of the most active volcanic regions in the world.

Glaciers and cooled lava beds each cover approximately one-tenth of the 39,768 square miles (103,000 square kilometres) of the country. The glaciers are a reminder of Iceland's close proximity to the Arctic Circle, which nearly touches its northernmost peninsula. The area covered by Vatna Glacier, the country's largest, is equal to the combined total area covered by all of the glaciers on the continent of Europe. The volcanoes, reaching deep into the unstable interior of the Earth, are explained by the fact that Iceland is located on the top of the Mid-Atlantic Ridge. It is estimated that since the year 1500 about one-third of the Earth's total lava flow has poured out of the volcanoes of Iceland.

Iceland was founded more than 1,000 years ago during the Viking Age of exploration. The capital, Reykjavík ("Bay of Smokes"), is near the site of the island's first farmstead. The early settlement, made up primarily of Norwegian seafarers and adventurers, fostered further excursions to Greenland and the coasts of North America, or Vinland. In spite of its physical isolation some 500 miles (800 kilometres) from Scotland, its nearest European neighbour, Iceland has remained throughout its history very much a part of European civilization. It is a Scandinavian country, modern in nearly every respect. The Icelandic sagas, most of which recount heroic episodes that took place at the time the island was settled, are regarded as among the finest literary achievements of the Middle Ages.

This article is divided into the following sections:

Physical and human geography 760
- The land 760
  - Relief
  - Drainage and soils
  - Climate
  - Plant and animal life
  - Settlement patterns
- The people 762
- The economy 763
  - Resources
  - Agriculture and fishing
  - Industry
  - Finance
  - Transportation
- Administration and social conditions 764
  - Government
  - Health and welfare
  - Education
- Cultural life 764

History 765
- Early history 765
  - Settlement (*c.* 870–*c.* 930)
  - Commonwealth (*c.* 930–1264)
- Iceland under foreign rule 765
  - Late Middle Ages (1264–*c.* 1550)
  - Growth of Danish royal power (*c.* 1550–*c.* 1830)
- Modern Iceland 766
  - Struggle for independence (*c.* 1830–1904)
  - Home rule and sovereignty (1904–44)
  - The Icelandic republic

Bibliography 767

Physical and human geography

THE LAND

Geologically young, Iceland contains about 200 volcanoes of various types. A new volcano erupting on the bottom of the sea between November 1963 and June 1967 created the island of Surtsey, off the southwestern coast. The new island grew to nearly one square mile in area and rose more than 560 feet (170 metres) above sea level, a total of 950 feet from the ocean floor.

Volcanism

Volcanic activity has been particularly frequent since the 1970s. A major eruption took place in 1973, when a volcano on the island of Heimaey spilled lava into the town of Vestmannaeyjar, an important fishing centre. Most of the 5,300 residents had to be evacuated, and, although the harbour remained intact, about a third of the town was destroyed. A number of eruptions have damaged geothermal projects and grazing land, but many eruptions have occurred outside populated areas.

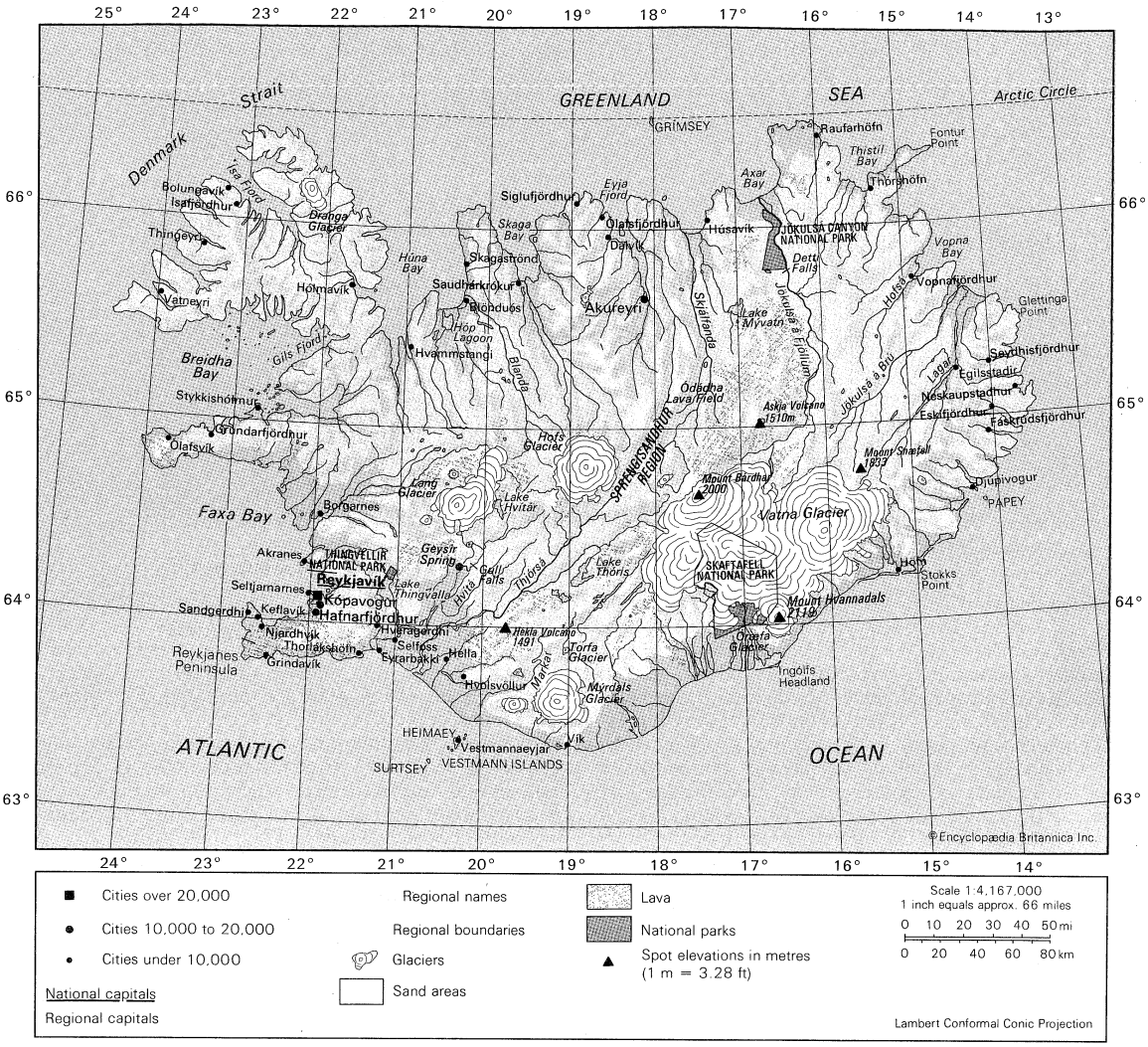
Relief. Iceland is largely a tableland broken up by structural faults. Its average elevation is 1,640 feet above sea level, but a quarter of the country lies at less than 650 feet. The highest point is at 6,952 feet (2,119 metres), atop Mount Hvannadals. The glaciers range in size from small ones in mountain recesses to enormous glacial caps topping extensive mountain ranges. Vatna Glacier covers an area of 3,240 square miles and is 3,000 feet deep at its thickest point.

Much of Iceland is underlain by basalt, a dark rock of igneous origin. None of this rock dates back farther than the Tertiary period, which began 66,400,000 years ago. The landscape in basaltic areas is one of plateau and fjord, characteristic of which are successive layers of lava visible one above the other on the valley sides. The basalt sheets tend to tilt somewhat toward the centre of the country. Glacial erosion played an important part in giving the valleys their present shape. The depressed zones between the basalt areas have extensive plateaus above which rise single volcanoes, table mountains, or other mountain masses with steep sides.

Iceland is richer in hot springs and solfataras—volcanic vents that emit hot gases and vapours—than any other country in the world. Alkaline hot springs are found in some 250 locations throughout the country. The largest, Deildartunguhver, emits 53 gallons (200 litres) of boiling water per second. The total power output of the largest of the 14 high-temperature solfatara regions, the Torfa Glacier area, is estimated to equal 1,500,000,000 watts. Earthquakes are frequent in Iceland but rarely result in serious damage. Most of the buildings erected since the mid-20th century have been designed and built of reinforced concrete to withstand severe earthquake shocks.

Traditional regions

Traditionally, Iceland has been divided according to the four points of the compass. The centre of the country is uninhabited. In the southwest, several good natural harbours have directed interest toward the sea, and the country's best fishing grounds lie off its shores. The southwest has little farmland because of its extensive lava fields and heaths. The middle west is divided between fishing and farming and has many places of great natural beauty. The western fjords have numerous well-sheltered harbours and good fishing grounds but little lowland suitable for agriculture. The north is divided into several smaller districts, each of which has relatively good farmland. The eastern fjords resemble the western fjords but have in addition an innerlying lowland. The southeast, locked between the glaciers and the sea, has a landscape of rugged splendour. The southern lowland comprises the main farming region. Soil and climatic conditions are favourable, and it is close to the country's largest market, Reykjavík and environs.



MAP INDEX

Political subdivisions

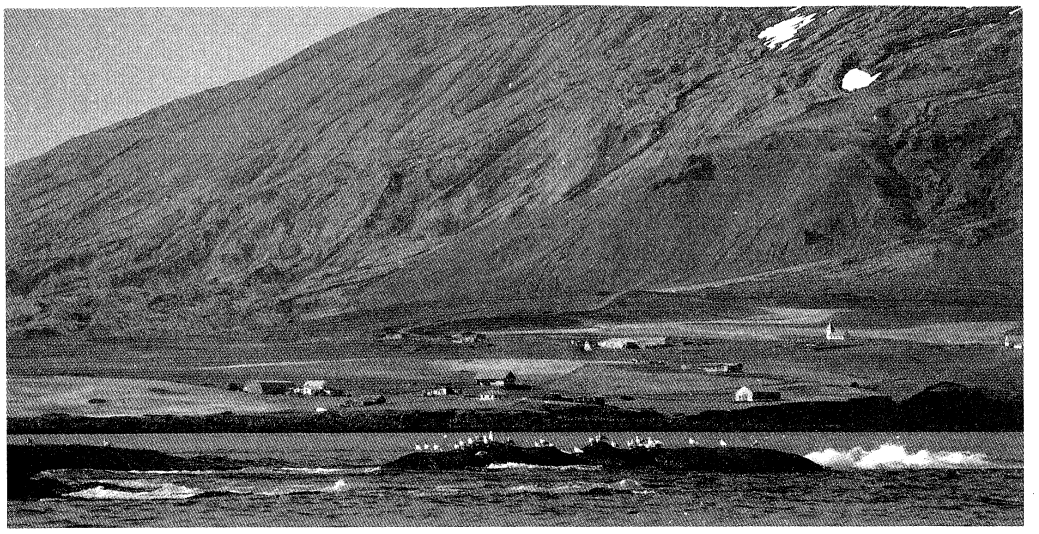
Austurland ..... 65 00 N 15 00 W
Höfudhborgarsvædhi ..... 64 15 N 21 30 W
Nordhurland Eystra ..... 65 30 N 17 00 W
Nordhurland Vestra ..... 65 15 N 19 45 W
Sudhurland ..... 64 15 N 19 30 W
Sudhurnes ..... 63 55 N 22 20 W
Vestfirdhir ..... 65 45 N 22 20 W
Vesturland ..... 64 45 N 21 30 W

Cities and towns

Akranes ..... 64 19 N 22 06 W
Akureyri ..... 65 40 N 18 06 W
Blönduós ..... 65 40 N 20 18 W
Bolungavík ..... 66 09 N 23 15 W
Borgarnes ..... 64 32 N 21 55 W
Dalvík ..... 65 58 N 18 32 W
Djúpivogur ..... 64 39 N 14 17 W
Egilsstadir ..... 65 16 N 14 25 W
Eskifjördhur ..... 65 04 N 14 00 W
Eyrarbakki ..... 63 52 N 21 09 W
Fáskrúdsfjördhur . 64 55 N 14 03 W
Grindavík ..... 63 50 N 22 26 W
Grundarfjördhur .. 64 55 N 23 15 W
Hafnarfjördhur ... 64 04 N 21 57 W
Hella ..... 63 50 N 20 24 W
Hofdhakaupstadhur, see Skagaströnd
Höfn ..... 64 15 N 15 13 W
Hólmavík ..... 65 43 N 21 41 W
Húsavík ..... 66 03 N 17 21 W
Hvammstangi ... 65 24 N 20 57 W
Hveragerdhi .... 64 00 N 21 12 W
Hvolsvöllur ..... 63 45 N 20 14 W
Innri-Njardhvík, see Njardhvík
Ísafjördhur ..... 66 05 N 23 09 W
Keflavík ..... 64 01 N 22 34 W
Kópavogur ..... 64 06 N 21 55 W
Neskaupstadhur . 65 09 N 13 42 W
Njardhvík (Innri-Njardhvík) . 63 58 N 22 31 W
Ólafsfjördhur .... 66 04 N 18 39 W
Ólafsvík ..... 64 53 N 23 43 W
Patreksfjördhur, see Vatneyri
Raufarhöfn ..... 66 27 N 15 57 W
Reykjavík ..... 64 09 N 21 57 W
Sandgerdhi ..... 64 03 N 22 42 W
Saudhárkrókur ... 65 45 N 19 39 W
Selfoss ..... 63 56 N 21 00 W
Seltjarnarnes .... 64 08 N 22 00 W
Seydhisfjördhur .. 65 16 N 14 00 W
Siglufjördhur .... 66 09 N 18 55 W
Skagaströnd (Hofdhakaupstadhur) ..... 65 50 N 20 19 W
Stykkishólmur ... 65 04 N 22 44 W
Thingeyri ..... 65 52 N 23 30 W
Thorlákshöfn .... 63 51 N 21 22 W
Thórshöfn ..... 66 12 N 15 20 W
Vatneyri (Patreksfjördhur) ..... 65 35 N 23 59 W
Vestmannaeyjar . 63 26 N 20 16 W
Vík ..... 63 25 N 19 01 W
Vopnafjördhur ... 65 45 N 14 50 W

Physical features and points of interest

Askja Volcano ... 65 03 N 16 48 W
Atlantic Ocean .. 63 15 N 19 00 W
Axar Bay ..... 66 15 N 16 45 W
Bárdhar, Mount .. 64 38 N 17 32 W
Blanda, *river* .... 65 39 N 20 18 W
Breidha Bay .... 65 15 N 23 15 W
Denmark Strait .. 66 30 N 24 00 W
Detti Falls ..... 65 49 N 16 24 W
Dranga Glacier .. 66 09 N 22 15 W
Eyja Fjord ..... 65 54 N 18 15 W
Faxa Bay ..... 64 24 N 23 00 W
Fontur Point .... 66 23 N 14 32 W
Geysir Spring ... 64 19 N 20 18 W
Gils Fjord ..... 65 26 N 21 50 W
Glettinga Point .. 65 30 N 13 37 W
Greenland Sea .. 66 30 N 16 00 W
Grímsey, *island* .. 66 33 N 18 00 W
Gull Falls ..... 64 20 N 20 08 W
Heimaey, *island* . 63 26 N 20 17 W
Hekla Volcano ... 64 00 N 19 40 W
Hofs Glacier .... 64 49 N 18 48 W
Hofsá, *river* ..... 65 43 N 14 48 W
Hóp Lagoon .... 65 32 N 20 30 W
Húna Bay ..... 65 50 N 20 50 W
Hvannadals, Mount ..... 64 01 N 16 41 W
Hvítá, *river* ..... 64 00 N 20 58 W
Hvítár, Lake .... 64 37 N 19 50 W
Ingólfs Headland . 63 48 N 16 39 W
Ísa Fjord ..... 66 10 N 23 00 W
Jökulsá á Brú, *river* ..... 65 40 N 14 16 W
Jökulsá á Fjöllum, *river* ... 66 02 N 16 27 W
Jökulsá Canyon National Park ... 66 00 N 16 30 W
Lagar, *river* ..... 65 40 N 14 18 W
Lang Glacier .... 64 39 N 20 12 W
Markar, *river* .... 63 32 N 20 05 W
Mýrdals Glacier .. 63 40 N 19 06 W
Mývatn, Lake ... 65 36 N 17 00 W
Ódádha Lava Field ..... 65 09 N 17 00 W
Öraefa Glacier ... 64 02 N 16 39 W
Papey, *island* ... 64 36 N 14 11 W
Reykjanes Peninsula ..... 63 50 N 22 41 W
Skaftafell National Park ... 64 15 N 17 00 W
Skaga Bay ..... 65 54 N 19 35 W
Skjálfanda, *river* . 65 59 N 17 38 W
Snæfell, Mount .. 64 48 N 15 34 W
Sprengisandur Region ..... 64 52 N 18 07 W
Stokks Point .... 64 14 N 14 58 W
Surtsey, *island* .. 63 18 N 20 36 W
Thingvalla, Lake . 64 11 N 21 09 W
Thingvellir National Park ... 64 16 N 21 05 W
Thistil Bay ..... 66 18 N 15 24 W
Thjórsá, *river* .... 63 47 N 20 48 W
Thóris, Lake .... 64 16 N 18 54 W
Torfa Glacier .... 63 54 N 19 00 W
Vatna Glacier ... 64 24 N 16 48 W
Vestmann (Westman) Islands ..... 63 25 N 20 18 W
Vopna Bay ..... 65 50 N 14 40 W



Arnarstapi, fishing village on Faxa Bay, Snæfell Peninsula, on the west coast of Iceland.
© Brian A. Vikander/West Light

Drainage and soils. Heavy rainfalls feed the numerous rivers and lakes in the glaciated landscape. Many of the lakes are dammed by lava flows or glacial ice. The presence of numerous waterfalls is typical of the young, geologically faulted landscape. The rivers are mainly debris-laden streams of glacial origin or clear streams formed by springs from underground water. In the regions not drained by glacier rivers (with their heavy loads of sediment), the coasts are irregular in outline, rocky, and incised with numerous fjords and smaller inlets. There are many fine natural harbours, since the innermost parts of the fjords often have been deepened by glacial erosion. Elsewhere, the coasts are regular, sandy, and lined extensively with offshore sandbars that form lagoons to the landward.

Iceland has soils of both mineral and organic composition. The mineral soils are basically a yellow-brown loess, formed by deposits of wind-transported matter. Both types of soil are suitable for agriculture, but, because of the slow rate of biological activity in the northern climate, they require heavy fertilization.

Climate. The climate of Iceland is maritime subarctic. It is influenced by the location of the country on the broad boundary between two contrasting air currents, one of polar and one of tropical origin. The climate is affected also by the confluence of two ocean currents, the Gulf Stream, from near the Equator, and the East Greenland Polar Current. The latter sometimes carries Arctic drift ice to Iceland's northern and eastern shores.

Seasonal fluctuations in temperature and precipitation depend chiefly on the weather fronts crossing the North Atlantic. The passage of a front south of Iceland brings relatively cold weather, especially in northern districts; one moving northeastward between Iceland and Greenland brings mild, rainy weather. Although its northernmost tips nearly touch the Arctic Circle, Iceland is much warmer than might be expected.

Mean annual temperatures for Reykjavík, on the west coast, for Akureyri, in the north, and for Vík, in the south, are, respectively, 41° F (5° C), 39° F (4° C), and 42° F (6° C). Mean January and July temperatures for the same locations are 31° F (0° C) and 52° F (11° C), 29° F (−2° C) and 52° F, and 34° F (1° C) and 52° F. In the northwest, snow falls on about 100 days a year; in the southeast, on about 40. Annual precipitation ranges from 16 inches (410 millimetres) on some high northern plateaus to more than 160 inches on the southern slopes of some ice-capped mountains. In the south it averages more than 80 inches. Gales are frequent, especially in winter, and occasionally heavy fogs may occur, but thunderstorms are rare. Reykjavík averages a bit more than 1,300 hours of bright sunshine a year. Often the aurora borealis appears, especially in fall and early winter.

Plant and animal life. Iceland lies on the border between a tundra vegetation zone of treeless plains and a zone of coniferous forests. Only about a quarter of the country is covered by a continuous carpet of vegetation. Bogs and moors are extensive, and sparse grasslands are often overgrazed. The remains of large birch forests are found in many places. A spruce reforestation program has had some success.

Foxes were the only land mammals in Iceland at the time of its settlement. Humans brought house and farm animals and accidentally introduced rats and mice. Later, reindeer were introduced, and many are still found in the northeastern highlands. After 1930, mink that were brought in for the production of furs also became wild in the country. Birdlife in Iceland is varied. Many nesting cliffs are densely inhabited, and the colony of ducks at Lake Mývatn, in the north, is the largest and most varied in the world. Salmon and trout abound in the lakes, brooks, and rivers. The fishing banks off the Icelandic shores are abundantly endowed with fish, although these resources have been considerably eroded by overexploitation by various nations following World War II. There are no reptiles or amphibians in Iceland.

Introduction of animal life

Settlement patterns. Until the end of the 19th century the population of Iceland was evenly distributed throughout the inhabitable parts of the country, since agriculture was the chief economic activity. With the advent of the fishing industry, commerce, and services at the beginning of the 20th century, the population has become increasingly concentrated in towns and villages. About 90 percent of the population now lives in communities of 200 or more people. The mainstay of most coastal towns is fishing and fish processing. The greatest population concentration is in Reykjavík and its environs, with about three-fifths of Iceland's total population. Reykjavík is a modern, cosmopolitan urban centre, which in addition to being the seat of government is the national focus of commerce, industry, higher education, and cultural activity. Akureyri, a fishing and industrial centre of about 14,000 inhabitants situated on the Eyja Fjord in the north, is second in importance. Keflavík is a fishing port near the major international airport that bears its name. The Vestmann (Westman) Islands, off the south coast, have some of the most important fishing and fish-processing industry in Iceland. Akranes, located across the bay from Reykjavík, is a service town for its region and has some industry. Ísafjördhur is a service town for the western fjord area. Siglufjördhur in the north was once the centre of the herring industry, but both the town and the fishing industry have declined in size and importance. Selfoss in the southern lowlands, serving the farming region, is the largest inland rural community in Iceland.

THE PEOPLE

The population of Iceland is extremely homogeneous. The inhabitants are descendants of settlers that began arriving

in AD 874 and continued in heavy influx for about 50 years thereafter. Historians differ on the exact origin and ethnic composition of the settlers but agree that between 60 and 80 percent were of Nordic stock from Norway. The remainder, from Scotland and Ireland, were largely of Celtic stock. The dominant language in the period of settlement was Old Norse, the language spoken throughout Scandinavia at that time. Through the centuries it has evolved into modern Icelandic, which is spoken throughout the country. There are no racial or ethnic distinctions. The early Nordic and Celtic stocks have long since merged, and the small number of subsequent immigrants have had no major effect on the population structure. The Lutheran faith has been the dominant religion since the mid-16th century. Almost the entire population belongs to the state-supported Evangelical Lutheran church. There is freedom of religion.

Homogeneity of people

Numbers and trends. The first comprehensive census in Iceland was taken in 1703, at which time 50,358 people were reported. The 18th century was marked by great economic hardship, and by 1801 the population had declined to 47,240. In the 19th century there began a slow increase, and by 1901 the count had risen to 78,470. Accelerated economic growth during the early decades of the 20th century was paralleled by a rapid growth in population, which in 1940 reached 121,474. During World War II and the early postwar period, there was rapid improvement in the standard of living and a new acceleration in the rate of population growth. The annual growth rate reached its highest peak during the 1950s; it has been declining since 1960, primarily because of a sharply reduced birth rate and continued emigration.

Migration. Between 1870 and 1901 there was a large-scale emigration to Canada and the United States because of unfavourable conditions in Iceland; during that period, emigrants outnumbered immigrants by the equivalent of about one-fifth of the 1901 population. Since 1901 emigration has continued to exceed immigration, although the net population loss through emigration has fluctuated. The percentage of the overall population lost annually through emigration, however, generally has declined.

THE ECONOMY

The Icelandic economy is based heavily on fishing and a broad variety of fish products, but it includes manufacturing and services as well. Exports account for nearly one-half of the gross national product. Despite the small population, the economy is modern, and the standard of living is on a par with that of other European countries.

Although most of Iceland's production is in private hands, government ownership has increased in certain areas. The government now owns many of the banking and financial institutions and, with local governments, most of the electricity-generating systems. Since World War II the government has aimed for full employment, a high rate of economic growth, protection of the balance of payments, price stability, and income equalization. The central government receives a major portion of its income from such indirect taxes as import duties and sales levies, whereas local governments derive most of their money from corporate and personal income taxes and from property levies.

Workers' associations

More than 90 percent of Iceland's labour force belongs to trade unions or employer organizations. The Icelandic Federation of Labour, a loose organization of almost all blue-collar and some white-collar workers, is the largest labour group. Farmers have a separate union, as do state and local government employees. Private employers are organized into the Employers' Federation. In general wage agreements are first negotiated between the Employers' Federation and the Icelandic Federation of Labour.

In the postwar period, fluctuations in fish catches and in foreign prices for fish products have produced fluctuating cycles in domestic income. Along with domestic pressures for higher wages and prices, this instability has caused persistent inflation and frequent currency devaluations. The productivity of the fishing industry, which has been higher than that of other export sources, has prevented any major diversification of exports and of manufacturing.

Resources. Iceland's energy resources are vast. Only about 10 percent of the potential hydroelectric power of the rivers has been tapped. Geothermal energy from the waters of the hot springs heats all of Reykjavík and several other communities; it provides steam for industrial energy and is used in commercial vegetable farming in greenhouses.

Agriculture and fishing. In 1930, 36 percent of Iceland's population lived on farms or were engaged in agriculture, but the proportion has fallen to only about 5 percent. The raising of livestock, mostly sheep, and dairy farming are the main occupations. About 20 percent of the land is arable, most of it used for grazing. The cultivation of grains is rare, whereas greenhouses are profitable businesses, especially in southern Iceland. Important agricultural products are livestock, hay, and grass. Iceland is virtually self-sufficient in fresh foods and dairy items, but it imports most other foodstuffs.

A steady improvement in Iceland's fishing technology has increased catches despite the gradual erosion of what once had been enormously rich fish populations off of the country's coasts. The concern over declining fish stocks has brought about several extensions of the fishing limit by the Icelandic government as well as increasingly strict catch quotas. Salmon farming grew rapidly in the 1980s. Cod and capelin make up about two-thirds of the total catch, and such white fish as cod and haddock are exported fresh, frozen, or in salted or dried form. The capelin and herring catches are usually reduced to oil and meal but are also salted. In all, fish and fish products make up about three-fourths of Iceland's exports. Such fishing-related industries as boat yards, repair docks, and net factories are also important.

Industry. In the early 1960s the government initiated a policy that encouraged foreign investment for the exploitation of the country's resources. A Swiss-owned aluminum smelter for imported ore began operation, and a plant operating with geothermal steam began the refining of diatomaceous earth mined from Lake Mývatn into pure diatomite for use in industrial filtration. Shell sand is mined for use in the cement industry, and perlite and pumice (volcanic cinders) are among the mineral resources. Increased attention has been given to extracting salt and minerals from seawater and processing seaweed for commercial purposes. There are also small industries that produce fertilizer, appliances, food and clothing, and books.

Tourism

Icelandair (Flugleidir), a major international carrier, has helped make the tourist trade increasingly important in the national economic picture. Since 1950, when some 4,000 tourists visited the island, the number of annual tourist visits has grown enormously, and Reykjavík has become an international convention centre.

Finance. Banking and other financial services are well developed. Iceland has several commercial banks with branches throughout the country and a number of savings banks. Other financial institutions include the government-owned investment credit funds, private pension funds, insurance companies, and the social insurance system. An important securities industry developed in the 1980s.

Transportation. The historic isolation of Iceland, caused by the rough seas of the North Atlantic and by the country's small market and industry, was broken when steam vessels began to visit Icelandic shores late in the 19th century. The first telegraph cable to Iceland was laid in 1906, and the Icelandic Steamship Company was founded in 1914. Before the 20th century, roads were practically unknown, the horse being the means of transportation throughout the island. Iceland has never built railroads, and in several places the Icelanders jumped straight from horseback to airplane.

Iceland's roads, as they now exist, mostly are unpaved, some being mere tracks. During the summer, driving is possible on the extensive sandy plains in the uninhabited interior, permitting expeditions between the glaciers. The Hringvegur ("Ring Road") stretches for 900 miles, forming a circle around the island.

The merchant-marine fleet carries most of Iceland's imports and exports. Icelandair and local service carriers are

important internally in compensating for the limited road system. Keflavík International Airport, the country's primary gateway, is located about 30 miles (50 kilometres) west of Reykjavík.

ADMINISTRATION AND SOCIAL CONDITIONS

Government. Iceland is a parliamentary democracy, with an elected president as head of state. In 1980 Vigdís Finnbogadóttir became the first woman elected president. The powers of the presidential office are very similar to those of other heads of state in western European democracies. A 63-member parliament, the Althing (Althingi), one of the oldest legislative assemblies in the world, is divided into upper and lower houses with equal power. Members serve for four years unless the parliament is dissolved. The executive branch is headed by a cabinet consisting of a prime minister and eight to 10 other ministers. The cabinet must have majority support in parliament, or at least be able to avoid censure, or it must resign. The judiciary consists of a supreme court and a system of lower courts, most of which hear both civil and criminal cases. Cases are heard and judged by appointed judges; there is no jury system. All lower court verdicts, except those rendered in cases involving certain misdemeanours, can be appealed to the supreme court. The constitution guarantees all citizens the civil rights and liberties customarily accorded in Western democracies.

Iceland is divided into eight regions (*landsvædi*) and more than 20 independent towns (*kaupstadhir*). There are also more than 200 rural communes (*hreppar*). Each town and commune administers local affairs through an elected council. The local government is chiefly responsible for education, municipal services, social affairs, and health services. Towns have the power to levy income taxes, property taxes, and a business turnover tax. Income taxes are their most important source of income.

The president, Althing, and local councils are elected every four years, but not necessarily all at once. All citizens 18 years of age and older may vote. There are six main political parties. The Independence Party, centre-to-conservative in political outlook, has commanded about 25 percent of the popular vote. The agrarian Progressive Party, with the next largest percentage of the popular vote, draws its strength from rural areas. The Social Democrats, People's Alliance, Citizen's Party, and Women's Alliance share the balance of the vote. The Citizen's Party is a splinter group from the Independence Party. The others are all left-of-centre.

Political parties

Iceland is a charter member of the United Nations and the North Atlantic Treaty Organization (NATO). In the postwar period it has centred its foreign policy around international cooperation and participated in joint Western defense efforts. It does not maintain an army or a navy. The United States, having assumed responsibility for Iceland's defenses, maintains a naval air station at Keflavík International Airport under NATO auspices.

Health and welfare. Iceland, with compulsory health insurance that finances many medical services, has one of the highest standards of public health in the world. Hospital services are provided entirely without charge, and other medical services are offered at very low cost. Dental care is partially subsidized for children three to five years of age and for the elderly. All contagious diseases that were major causes of death in the 19th century have been virtually eradicated. Mortality due to such epidemic diseases as influenza and measles has been minimized. Heart disease and cancer together account for about one-half of all deaths. Welfare services include unemployment insurance, old-age and disability pensions, family allowances, childbearing cash grants, and sickness benefits. The medical and welfare systems are financed by central and local government.

Education. Almost all schools from the primary level through the university are free. Education is compulsory through age 16, and secondary and higher education is widely available. Students can enroll in four-year academic colleges at the age of 15 or 16. Graduation from one of these colleges entitles the student to admission to the University of Iceland, founded in 1911, in Reykjavík. There are also a number of technical, vocational, and specialized schools.

CULTURAL LIFE

Icelanders are proof that a rich cultural life can be developed despite a small population. For centuries Iceland's writers concentrated on rewriting the great sagas of the Middle Ages. Other arts included weaving, silver crafting, and wood carving. In the 20th century, poetry declined in popularity, but writing was pursued with eagerness, as were painting and theatre, both relatively new traditions in the country. The government heavily subsidizes museums, and the National Theatre in Reykjavík provides annual stipends to playwrights.

The Reykjavík area, which supports two professional theatres, a symphony orchestra, an opera, and several art galleries, bookstores, cinemas, and museums, has a cultural activity that compares favourably with towns or cities several times its size. Reykjavík's location between Europe and North America brings many celebrities, especially in music, and various exhibitions stop there for a day or a week on their way between continents. Biennial international art festivals are held in Reykjavík.

Formerly, art in Iceland was often connected with religion, first with the Roman Catholic church and later the Lutheran. The first professional secular painters appeared in Iceland about 1800. Gradually increasing in number, the painters seemed eager to portray the character and beauty of their country. The best known was Jóhannes Kjarval. There is much activity in painting in Iceland, where a number of artists have fused foreign influences with a local heritage. The old traditions in silver working have been retained, the most characteristic of which is the use of silver thread for ornamentation. Newer trends also have flourished.

The literary tradition of the saga was revived in the mid-19th century, and the old tales were increasingly elaborated. The writers used Icelandic settings almost exclusively. Several writers became well known abroad, including Halldór Laxness, who received the Nobel Prize for Literature in 1955. Some writers have written for the theatre, and their work has been more international in theme and setting.

Music also has been cultivated in Iceland. The programs of the Icelandic Symphony Orchestra are drawn from an international repertoire. One or more operas or musicals are performed every year at the National Theatre and at the Reykjavík Opera. Few Icelandic composers are known outside the country. On the lighter side, popular music is enjoyed in all its rapidly changing styles as it is throughout the West. The Music Society operates a college of music.

Traditional folk crafts

National folk traditions in applied art have been given a new strength by a lively trade in art and by souvenir shops. Many people in the country take part in this industry, creating high-quality goods. For most residents, it is a hobby that produces extra income. Old designs and forms have been revived, some modified to please modern tastes. Wool, knitted or woven, is the most commonly used material.

In sports a special old form of wrestling is maintained, but pony trekking is much more popular. Swimming, in naturally heated swimming pools, and various ball games are also popular. Chess approaches the status of a national sport, and Iceland frequently hosts international chess tournaments. Seasonal sports include fishing and hiking in summer and skiing in winter.

The National Library of Iceland was founded in 1818 and the University Library in 1940. The National Archives were founded in 1882. The National Museum of Iceland, dating from 1863, has collections representing native Icelandic culture dating from the Viking Age; many old houses and ruins throughout the country are preserved under its auspices.

The Art Gallery of Iceland was founded in 1885. The great majority of its works are by modern Icelandic artists. The Natural History Museum, founded in 1889, has always been small as a museum, but its departments of geology and geography, botany, and zoology have been active in various areas of research.

The National Theatre began operation in 1950. It performs Icelandic as well as foreign plays, operas, ballets, and musicals. It sponsors some activities outside the theatre in Reykjavík. Some book clubs are active, the oldest one being the Icelandic Literary Society, founded in 1816.

(Va.K./B.M.)

For statistical data on the land and people of Iceland, see the *Britannica World Data* section in the BRITANNICA WORLD DATA ANNUAL.

History

EARLY HISTORY

Settlement (c. 870–c. 930). Iceland apparently has no prehistory. According to stories written down some 250 years after the event, the country was discovered and settled by Norse people in the Viking Age. The oldest source, *Íslendingabók* (*The Book of the Icelanders*), written in about 1130, sets the period of settlement at about AD 870–930. The other main source, *Landnámabók* (*The Book of Settlements*), of 12th-century origin but known only in later versions, states explicitly that the first permanent settlers, Ingólfr Arnarson and his wife, Hallveig Fródadóttir, came to Iceland to settle in the year 874. They came from Norway and chose as their homestead a site that Ingólfr named Reykjavík. *The Book of Settlements* then enumerates more than 400 settlers who sailed with their families, servants, and slaves to Iceland to stake claims to land there. Most of the settlers came from Norway, but some came from other Nordic countries and from the Norse Viking Age settlements in the British Isles.

Sources

Some scholars have found it difficult to believe that such a large and habitable country remained unpopulated until the 9th century. On the whole, however, archaeological finds support the documentary evidence and place Iceland among Norse Viking Age settlements of the late 9th or early 10th century. The Icelandic language testifies to the same origin: Icelandic is a Nordic language and is most closely related to the dialects of western Norway.

Although the island was not populated until the Viking Age, Iceland probably had been known to people long before that time. The 4th-century-BC Greek explorer Pytheas of Massalia (Marseille) described a northern country that he called Thule, located six days' sailing distance north of Britain. In the 8th century, Irish hermits who had begun to sail to Iceland in search of solitude also called the island Thule. It is unknown, however, if Pytheas and the hermits were describing the same island. According to the early Icelandic sources, some Irish monks were living in Iceland when the Nordic settlers arrived, but these monks soon left because they were unwilling to share the country with heathens.

Commonwealth (c. 930–1264). At the time of Iceland's settlement, Norse people worshiped gods whom they called *æsir* (singular *áss*), and this religion has left behind an extensive mythology in Icelandic literature. Thor seems to have been the most popular of the pagan gods in Iceland, although Odin is thought to have been the highest in rank. It appears that heathen worship was organized around a distinct class of priest-chieftains, called *godar* (singular *godi*), of which there were about 40. In the absence of royal power in Iceland, the *godar* were to form the ruling class in the country.

The *godar*

By the end of the settlement period, a general Icelandic assembly, called Althing, had been established and was held at midsummer on a site that came to be called Thingvellir. This assembly consisted of a law council (*lögrétta*), in which the *godar* made and amended the laws, and a system of courts of justice, in which householders, nominated by the *godar,* acted on the panels of judges. At the local level, three *godar* usually held a joint assembly in late spring, at which a local court operated, again with judges nominated by the *godar.* All farmers were legally obliged to belong to a chieftaincy (*godord*) but theoretically were free to change their allegiance from one *godi* to another; the *godar* were allotted a corresponding right to expel a follower. Some scholars have seen in this arrangement a resemblance to the franchise in modern societies. On the other hand, there was no central authority to ensure that the farmers would be able to exercise their right in a democratic way. No one was vested with executive power over the country as a whole. In any case, no trace of democratic practice reached farther down the social scale than to the heads of farming households: women and workers (either free or enslaved) had no role in the political system.

Christianization. By the end of the 10th century the Norwegians were forced by their king, Olaf I Tryggvason, to accept Christianity. The king also sent missionaries to Iceland, who according to 12th-century sources were highly successful in converting the Icelanders. In 999 or 1000 the Althing made a peaceful decision that all Icelanders should become Christians. In spite of this decision, the *godar* retained their political role, and many of them probably built their own churches. Some were ordained, and as a group they seem to have closely controlled the organization of the new religion. Two bishoprics were established, one at Skálholt in 1056 and the other at Hólar in 1106. Literate Christian culture also transformed lay life. Codification of the law was begun in 1117–18. A little later the Icelanders began to write their sagas, which were to reach their pinnacle of literary achievement in the next century.

Economic life. Historians believe that early Icelandic society was a prosperous one. The country proved to be well suited for sheep and cattle, and both were raised for meat and milk. The sheep also yielded wool, and homespun cloth became the chief export. There was some agriculture, but grain was always imported. Timber was also imported: the only indigenous wood was dwarf birch. However abundant driftwood may have been, it could not satisfy the needs of the whole population. The Icelanders built large turf-clad houses on bulky timber frames, and some of the churches were built entirely of timber.

In spite of the seeming abundance, the end was coming for an independent Icelandic commonwealth. In Norway, royal power gained strength in the early 13th century, and the king set himself the aim of uniting all Norwegian Viking Age settlements under his reign. By that time, a score of powerful *godar,* belonging to some five families, held almost all the chieftaincies in Iceland. The mid-13th century is characterized by a bloody struggle for power among these chieftains. Finally, in 1262–64 all Icelandic chieftains and representatives of the farmers were persuaded to swear allegiance to the king of Norway, partly in the hope that he would bring peace to the country.

End of the commonwealth

ICELAND UNDER FOREIGN RULE

Late Middle Ages (1264–c. 1550). To a large extent, Iceland was ruled separately from Norway. It had its own law code, and the Althing continued to be held at Thingvellir, though mainly as a court of justice. Most of the royal officials who succeeded the chieftains were Icelanders. In 1380 the Norwegian monarchy entered into a union with the Danish crown, but that change did not affect Iceland's status within the realm as a personal *skattland* ("tax land") of the crown.

Economic growth and decline. A fundamental change in Iceland's economy took place in the early 14th century when Norwegian merchants began to import dried fish from Iceland to Bergen, Nor. English merchants in Bergen became acquainted with Icelandic fish supplies, and shortly after 1400 they themselves began sailing to Iceland and either catching fish or buying it from local fishermen. The Danish crown repeatedly tried to stop English trade in Iceland but lacked the naval power with which to defend its remote possession. One of the royal governors was killed by the English when he tried to stop their trade, an event that led indirectly to clashes between Denmark and England (1468–73). In the early 16th century, English interest in Iceland declined, partly because rich fishing grounds had been discovered off the North American coast of Newfoundland. Instead, Germans became the chief foreigners to fish and trade in Iceland.

Importance of fishing

In spite of the rise of a profitable export industry, it is generally believed that Iceland's economy deteriorated in the late Middle Ages. The birchwood that had covered great parts of the country was gradually depleted, in part

because it was excellent for making charcoal. The destruction of the woodland, together with heavy grazing, led to extensive soil erosion. The climate also became more severe, and grain growing was given up altogether. At the same time, more and more of the land was acquired by ecclesiastical institutions and wealthy individuals, to whom the farmers had to pay rent.

The Reformation. The Lutheran Reformation, which was instituted in Denmark in the 1530s, met greater resistance in Iceland than anywhere else in the realm. In 1541 the bishop of Skálholt was captured by the governor, and Lutheranism was introduced in his diocese, which covered three-quarters of the country. In the northern diocese of Hólar, Bishop Jón Arason was to hold out against Lutheranism for a decade longer. In 1550 he was finally captured and beheaded, without benefit of law or clergy, and all resistance to the Reformation ended. Jón Arason's death is traditionally taken to mark the end of the Middle Ages in Iceland.

Growth of Danish royal power (c. 1550–c. 1830). After the Reformation the royal treasury confiscated all lands that had belonged to the Icelandic monasteries. Also during the 16th century the German traders were ousted. In 1602 all foreign trade in Iceland was monopolized by a royal decree and handed over to Danish merchants, who paid a rent on it to the crown. This arrangement lasted intact until 1787, when the monopoly was abolished. Only subjects of the Danish crown, however, were permitted to carry on foreign trade, a restriction that remained in force until 1855.

Danish monopoly of trade

On the constitutional level, also, the Danish crown increased its hold on Iceland, at least in formal terms. In 1661 Frederick III introduced an absolute monarchy in Denmark and Norway, and in the following year his absolutism was acknowledged in Iceland. This event was not of any great immediate significance in Iceland; local officials, most of whom were Icelanders, continued to make important political decisions. Danish officials in Copenhagen rarely had enough knowledge of or interest in Icelandic affairs to enforce their will if the Icelandic officials in Iceland were unanimous on a different policy.

Nevertheless, the bureaucratic state, which formed the backbone of absolutism, was gradually introduced into Iceland. An essential part of that development was the emergence of a town nucleus in Reykjavík, the first one in this hitherto entirely rural country. In the 1750s a tiny village grew up in Reykjavík as a result of a semiofficial attempt to start a wool-processing factory there. Within half a century the two ancient bishoprics were united, with the bishop residing in Reykjavík. The Althing was abolished in 1800, and an appeals court was set up in Reykjavík to succeed it. A few years later the Danish governor also settled in the town, which by then had about 300 inhabitants.

Economic conditions

Only since the 18th century has there been sufficient information available to allow for a reliable survey of Iceland's economy. The first census, for example, was held in 1703, revealing that the population was just over 50,000. The main occupation of most of the inhabitants was farming, but an important auxiliary occupation, undertaken mostly by rural labourers on the southern and western coasts in late winter and spring, was fishing. With few exceptions, labourers were obliged to stay in the domestic service of a farmer, and the establishment of permanent households in fishing stations was severely restricted. Thus, the landowners, with most of the native officials in their number, succeeded in monopolizing fishing, and they prevented it from becoming an independent industry.

The 18th century, a time of prosperity elsewhere in the Danish realm, was a period of decline and increasing poverty in Iceland. In the 1780s the country was plagued by famine—caused by a volcanic eruption and subsequent years of cold weather—which killed one-fifth of the population. In spite of these hardships, it appears that few Icelanders were critical of the status of the country within the Danish realm. In 1809 a Danish adventurer, Jørgen Jørgensen, seized power in Iceland for two months. When he was removed and Danish power was restored, he received no support from the Icelandic population. Five years later, when Norway was united politically with Sweden, no Icelander expressed a wish to follow suit.

MODERN ICELAND

Struggle for independence (c. 1830–1904). In the 1830s the Danish crown eased its absolutism by setting up four consultative diets in the realm. Iceland was allotted two representatives in one of them, the assembly of the Danish Isles meeting at Roskilde, Den. This arrangement was never popular in Iceland, and the adherents to emerging nationalist and liberal sentiments wanted the Icelandic Althing to be restored on its ancient site as a consultative assembly for the nation. In 1840 Christian VIII decided to grant them their wish, and five years later a restored Althing, vested with consultative power, met for the first time, not at Thingvellir, however, as originally intended, but in Reykjavík. Franchise to the assembly was almost entirely restricted to officials and (male) farmers.

In 1848, only three years after the first convention of the Althing, Frederick VII renounced his absolute power, and a constitutional assembly was summoned to prepare a representative democracy in Denmark. This led inevitably to the question of what was to become of Iceland in the new form of government. By this time, Iceland had a relatively undisputed political leader, Jón Sigurdhsson, a philologist living in Copenhagen. Sigurdhsson argued that the king could only give his absolute rule over Iceland back to the Icelanders themselves, since they were the ones who had surrendered it to him in 1662. This claim was met with a royal pledge that the constitutional status of Iceland would not be decided until the Icelanders had discussed the matter at a special assembly in Iceland. This assembly met in 1851, but no agreement could be reached between the Icelandic representatives and the Danish government. The assembly was dissolved in disappointment. A stalemate of more than 20 years ensued, but the Althing decided to use the occasion of the millennium of Iceland's settlement (1874) to accept the status that Danish authorities were by then willing to grant. Thus, in 1874 the king presented Iceland with a constitution whereby the Althing was vested with legislative power in internal affairs. As before, however, the cabinet minister responsible for Iceland was the minister of justice in the Danish government.

Constitution of 1874

For an additional three decades the Icelanders fought to get executive power transferred to Iceland. In 1901 the path was opened when rule by parliamentary majority was introduced in Denmark and the Liberal Party—always more positive than the Conservatives toward the Icelanders—came into power. In 1904 Iceland got home rule, and the first Icelandic minister opened his office in Reykjavík. At the same time, rule by parliamentary majority was introduced.

The high level of political activity in 19th-century Iceland stands in sharp contrast with its economic stagnation. The considerable growth of Iceland's population put increasing strain on the badly eroded rural areas, and for many people the only visible solution was emigration to North America. Some 15,000 Icelanders emigrated between 1870 and 1914, most of them to Canada. Virtually the only successful technical innovation during that period was the introduction of decked fishing vessels, which made it possible to catch fish farther offshore than could be done on open boats. Still, at the beginning of the 20th century, more than half of the annual catch was still taken in open boats.

Home rule and sovereignty (1904–44). The period of home rule (1904–18) was one of rapid progress. Motors were installed in many of the open fishing boats, and a number of steam-driven trawlers were acquired. The country was connected by telephone cable with Europe. School attendance was made compulsory for children in towns and villages, and a number of schools were built. The University of Iceland was established in Reykjavík, which by 1918 had a population of 15,000. All restrictions on the freedom to move to the fishing villages were either abolished or quietly forgotten. There was a radical transformation in the occupational structure of the country, which, in turn, led to the advent of a labour movement. In 1916 a national organization of trade unions was es-

tablished. By then, unions were already widely accepted by employers as negotiating bodies, but their formal status was not legalized until 1938. In the political arena, democracy was extended to new groups in the society. Women and propertyless men were given the franchise, subject to certain qualifications, in 1915. Four years earlier, a law had been passed that gave women the right to attend schools of higher education, enter into the professions, and occupy any public office in the country.

The union

The struggle continued for greater autonomy until the dispute with Denmark was solved by an agreement. On Dec. 1, 1918, Iceland became a separate state under the Danish crown, with only foreign affairs remaining under Danish control. Either party, however, had the right to terminate the treaty after 25 years if negotiations about its renewal proved fruitless.

The struggle for independence that had shaped Icelandic politics for almost a century now subsided, and in the 1920s a new system of political parties based on class divisions emerged. Class antagonism grew more severe during the Great Depression of the 1930s; the depression was prolonged in Iceland when the outbreak of the Spanish Civil War in 1936 closed the important Spanish market for Icelandic fish. The problem of high unemployment persisted until after the outbreak of World War II.

The German occupation of Denmark in April 1940 effectively dissolved the union between Iceland and Denmark. A month later, British forces occupied Iceland. In 1941 the United States took over the defense of Iceland and stationed a force of 60,000 in the country. The foreign forces brought employment, prosperity, and high inflation to the population, which then numbered about 120,000.

Because of the war situation, it was not possible for Iceland and Denmark to renegotiate their treaty. In spite of great resentment in Denmark, the Icelanders decided to terminate the treaty, break all constitutional ties with Denmark, and establish a republic. On June 17, 1944, now celebrated as National Day, the Icelandic republic was founded at Thingvellir, with Sveinn Björnsson as its first president.

The Icelandic republic. Since the prosperous years of World War II, Iceland has developed into a modern welfare state with growing production and consumption. A rapid restoration of the trawler fleet after the war prevented the return of prewar unemployment. Fish freezing became a highly technical industry and the mainstay of Iceland's exports. The economy became characterized by expansion, full employment, high inflation, and much unprofitable investment. It became normal to work overtime and for women to enter the labour market.

The tendency toward overexpansion is partly due to weak political leadership: no party has ever held an absolute majority in the Althing, and generally the country has been ruled by a coalition government. The only coalition to have lasted more than one electoral term without interruption was that formed by the conservative Independence and the Social Democratic parties, which remained in power for three terms (1959–71). The two other parties that have shown a preference to form coalitions have been the agrarian/liberal Progressive Party and the left-wing socialist People's Alliance.

Political coalition

This tendency in Icelandic politics to blur the political left and right has been continued in other coalitions: the Independence and the Social Democratic parties, for example, have shared a friendlier attitude toward cooperation with other Western countries, while the Progressives and the People's Alliance have been more wary of risking the loss of national autonomy and culture. This contrast has come to a head particularly in controversies over the two recurrent issues in postwar politics: defense and the extension of fishing limits.

Defense. When Iceland joined the North Atlantic Treaty Organization (NATO) in 1949—and subsequently received American forces in 1951—the Socialist Party, which was absorbed into the People's Alliance in 1956, was the only party unanimously opposed to the policy. Since the end of the war, the main dividing line in foreign policy has been between the Socialist Party/People's Alliance, which demands the restoration of neutrality, and the three supporters of NATO. The Progressives, however, have repeatedly, although unsuccessfully, united with the People's Alliance with the aim of expelling the Americans. On one occasion, the Labour Party reluctantly entered a coalition that had this policy on its program. The Independence Party has always been a staunch supporter of the American-manned base in the country.

Fishing limits. Since the war, Iceland has gradually extended its fishing limits from three nautical miles (six kilometres) in 1950 to 200 miles in 1975. This extension provoked strong protests from the United Kingdom and West Germany, and the British Navy was repeatedly sent to the Icelandic fishing grounds to protect British trawlers. The struggle with Britain, commonly known as the "cod wars," came to an end in 1976, when Britain recognized the 200-mile limit. Although all of the political parties supported the claim for Iceland's dominance over the fishing grounds, only the more "nationalistic" parties were willing to risk Iceland's good relations with its NATO partners.

Extension of fishing limits

(Gu.K.)

For later developments in the history of Iceland, see the *Britannica Book of the Year* section in the BRITANNICA WORLD DATA ANNUAL.

For coverage of related topics in the *Macropædia* and *Micropædia,* see the *Propædia,* sections 923 and 972.

BIBLIOGRAPHY. *Physical and human geography:* JÓHANNES NORDAL and VALDIMAR KRISTINSSON (eds.), *Iceland 1986: Handbook* (1987), provides a comprehensive general survey. Geological, geothermal, geographic, and ecological characteristics of the country are studied in THOMAS F.W. BARTH, *Volcanic Geology, Hot Springs, and Geysers of Iceland* (1950); SIGURDUR THORARINSSON, *Hekla: A Notorious Volcano,* trans. from Icelandic (1970); STURLA FRIDRIKSSON, *Surtsey* (1975); R.N. STEWART, *Rivers of Iceland* (1950); HUBERTUS PREUSSER, *The Landscapes of Iceland* (1976); BJÖRN JÓHANNESSON, *The Soils of Iceland* (1960); and STEINDÓR STEINDÓRSSON FRÁ HLÖDUM, *Studies on the Mire-Vegetation of Iceland* (1975). Illustrated descriptions and guidebooks include HJÁLMAR R. BÁRDARSON, *Ice and Fire: Contrasts of Icelandic Nature,* trans. from Icelandic, 3rd ed. (1980); MAX SCHMID, *Iceland: The Exotic North* (1985); STEINDÓR STEINDÓRSSON FRÁ HLÖDUM (ed.), *Iceland Road Guide,* 3rd updated ed. (1981; originally published in Icelandic, 1974); BJARNI GUDMUNDDSSON and THORLEIFUR THÓRDARSON, *Iceland,* 3rd rev. ed. (1981); and DAVID WILLIAMS, *Iceland: The Visitor's Guide* (1985).

Icelandic society, culture, and political orientation are reviewed in RICHARD F. TOMASSON, *Iceland, the First New Society* (1980); HJÁLMAR R. BÁRDARSON, *Iceland: A Portrait of Its Land and People* (1982; originally published in Icelandic, 1965); KRISTJÁN ELDJÁRN, *Ancient Icelandic Art* (1957); CAROL J. CLOVER and JOHN LINDOW (eds.), *Old Norse-Icelandic Literature: A Critical Guide* (1985); and BENEDIKT GRÖNDAL, *Iceland: From Neutrality to NATO Membership* (1971).

History: KNUT GJERSET, *History of Iceland* (1924), is still the most extensive survey of the country up to the 20th century. SIGURDUR A. MAGNÚSSON, *Northern Sphinx: Iceland and Icelanders from the Settlement to the Present* (1977), deals with the commonwealth period and the events of the 20th century but treats very little of the intermediate period. Scholarly studies include JÓN JÓHANNESSON, *A History of the Old Icelandic Commonwealth: Íslendinga Saga,* trans. from Icelandic (1974); and JESSE L. BYOCK, *Medieval Iceland: Society, Sagas, and Power* (1988). For this century, see MAGNÚS S. MAGNÚSSON, *Iceland in Transition: Labour and Socio-Economic Change Before 1940* (1985); and GYLFI TH. GÍSLASON, *The Problem of Being an Icelander,* trans. from Icelandic (1973), a short survey with an emphasis on the postwar period. BJÖRN THORSTEINSSON, *Island,* trans. from Icelandic (1985), is a historical survey in Danish. SIGURDUR LÍNDAL (ed.), *Saga Íslands* (1974–), is a comprehensive multivolume history in Icelandic, still in progress.

(B.M./Gu.K.)

Ideology

An ideology is a form of social or political philosophy in which practical elements are as prominent as theoretical ones; it is a system of ideas that aspires both to explain the world and to change it.

This article describes the nature, history, and significance of ideologies in terms of the philosophical, political, and international contexts in which they have arisen. For discussions of particular categories of ideology, see the *Macropædia* article SOCIO-ECONOMIC DOCTRINES AND REFORM MOVEMENTS, MODERN. Specific ideologies are described in the *Micropædia* under such headings as COMMUNISM and FASCISM.

The article is divided into the following sections:

Origins and characteristics of ideology 768
The philosophical context 768
  Ideology and religion
  Ideology in early political philosophy
  Hegel and Marx
  The sociology of knowledge
The political context 770
  Ideology, rationalism, and romanticism
  Ideology and terror
  Ideology and pragmatism
The context of international relations 771
  Ideology in the World Wars
  Ideology of the Cold War
Bibliography 772

ORIGINS AND CHARACTERISTICS OF IDEOLOGY

Science of ideas

The word first made its appearance in French as *idéologie* at the time of the French Revolution, when it was introduced by a philosopher, A.-L.-C. Destutt de Tracy, as a short name for what he called his "science of ideas," which he claimed to have adapted from the epistemology of the philosophers John Locke and Étienne Bonnot de Condillac, for whom all human knowledge was knowledge of ideas. The fact is, however, that he owed rather more to the English philosopher Francis Bacon, whom he revered no less than did the earlier French philosophers of the Enlightenment. It was Bacon who had proclaimed that the destiny of science was not only to enlarge man's knowledge but also to "improve the life of men on earth," and it was this same union of the programmatic with the intellectual that distinguished Destutt de Tracy's *idéologie* from those theories, systems, or philosophies that were essentially explanatory. The science of ideas was a science with a mission; it aimed at serving men, even saving them, by ridding their minds of prejudice and preparing them for the sovereignty of reason.

Destutt de Tracy and his fellow *idéologues* devised a system of national education that they believed would transform France into a rational and scientific society. Their teaching combined a fervent belief in individual liberty with an elaborate program of state planning, and for a short time under the Directory (1795–99) it became the official doctrine of the French Republic. Napoleon at first supported Destutt de Tracy and his friends, but he soon turned against them, and in December 1812 he even went so far as to attribute blame for France's military defeats to the influence of the *idéologues,* of whom he spoke with scorn.

Thus *ideology* has been from its inception a word with a marked emotive content, though Destutt de Tracy presumably had intended it to be a dry, technical term. Such was his own passionate attachment to the science of ideas, and such was the high moral worth and purpose he assigned to it, that the word *idéologie* was bound to possess for him a strongly laudatory character. And equally, when Napoleon linked the name of *idéologie* with what he had come to regard as the most detestable elements in Revolutionary thought, he invested the same word with all of his feelings of disapprobation and mistrust. Ideology was, from this time on, to play this double role of a term both laudatory and abusive not only in French but also in German, English, Italian, and all the other languages of the world into which it was either translated or transliterated.

Some historians of philosophy have called the 19th century the age of ideology, not because the word itself was then so widely used, but because so much of the thought of the time can be distinguished from that prevailing in the previous centuries by features that would now be called ideological. Even so, there is a limit to the extent to which one can speak today of an agreed use of the word. The subject of ideology is a controversial one, and it is arguable that at least some part of this controversy derives from disagreement as to the definition of the word *ideology.* One can, however, discern both a strict and a loose way of using it. In the loose sense of the word, ideology may mean any kind of action-oriented theory or any attempt to approach politics in the light of a system of ideas. Ideology in the stricter sense stays fairly close to Destutt de Tracy's original conception and may be identified by five characteristics: (1) it contains an explanatory theory of a more or less comprehensive kind about human experience and the external world; (2) it sets out a program, in generalized and abstract terms, of social and political organization; (3) it conceives the realization of this program as entailing a struggle; (4) it seeks not merely to persuade but to recruit loyal adherents, demanding what is sometimes called commitment; (5) it addresses a wide public but may tend to confer some special role of leadership on intellectuals. In this article the noun *ideology* is used only in its strict sense; the adjective *ideological* is used to refer to ideology as broadly defined.

Ideology loosely and strictly defined

On the basis of the five features above, then, one can recognize as ideologies systems as diverse as Destutt de Tracy's own science of ideas, the Positivism of the French philosopher Auguste Comte, Communism and several other types of Socialism, Fascism, Nazism, and certain kinds of nationalism. That all these "-isms" belong to the 19th or 20th century may suggest that ideologies are no older than the word itself—that they belong essentially to a period in which secular belief has increasingly replaced traditional religious faith.

THE PHILOSOPHICAL CONTEXT

Ideology and religion. Ideologies, in fact, are sometimes spoken of as if they belonged to the same logical category as religions. Both are assuredly in a certain sense "total" systems, concerned at the same time with questions of truth and questions of conduct; but the differences between ideologies and religions are perhaps more important than the similarities. A religious theory of reality is constructed in terms of a divine order and is seldom, like that of the ideologist, centred on this world alone. A religion may present a vision of a just society, but it cannot easily have a practical political program. The emphasis of religion is on faith and worship; its appeal is to inwardness and its aim the redemption or purification of the human spirit. An ideology speaks to the group, the nation, or the class. Some religions acknowledge their debt to revelation, whereas ideology always believes, however mistakenly, that it lives by reason alone. Both, it may be said, demand commitment, but it may be doubted whether commitment has ever been a marked feature of those religions into which a believer is inducted in infancy.

Even so, it is in certain religious movements that the first ideological elements in the modern world can be seen. The city of Florence, which in so many fields witnessed

Savonarola as an ideologist

the birth of modernity, produced perhaps the first "ideological" Christian. The attempt of Girolamo Savonarola to construct a puritan utopia was marked by several of the qualities by which one recognizes a modern ideology: Savonarola treated the vision of a Christian community as a model that men should actually seek to realize in the here and now. His method was to dominate the state through an appeal to the populace, and then to use the powers of the state to control both the economy and the private lives of the citizens. The enterprise was given a militant spirit; it was presented by Savonarola as being at one and the same time an outward struggle against papal corruption, the commercial ethos, and Renaissance Humanism, and an inward struggle against worldly ambitions and carnal desires.

Savonarola had numerous followers in his attempt to give Christianity an ideological dimension: he inspired Calvin's Geneva and the Puritan communities of the New World. Indeed, in both the Reformation and the Counter-Reformation, when Christianity was invested with a new militàncy and a new intolerance, when a new emphasis was placed on creeds and conversion, religion itself moved that much nearer to ideology.

Ideology in early political philosophy. The Italian political philosopher Niccolò Machiavelli was one of Savonarola's sharpest critics, but he was also, like him, a precursor of modern ideologists. Historians who speak of him only as an immoralist overlook the extent to which Machiavelli was a man with an ideal—a republican ideal. Rousseau recognized this when he spoke of *The Prince* as a "handbook for republicans." Machiavelli's dream was to see revived in modern Italy a republic as glorious as that of ancient Rome, and he suggested that it could be achieved only by means of a revolution that had the strength of will to liquidate its enemies. Machiavelli was the first to link ideology with terror, but he was too much of a political scientist to enact the role of the ideologue.

Ideology in 17th-century England

Seventeenth-century England occupies an important place in the history of ideology. Although there were then no fully fledged ideologies in the strict sense of the term, political theory, like politics itself, began to acquire certain ideological characteristics. The swift movement of revolutionary forces throughout the 17th century created a demand for theories to explain and justify the radical action that was often taken. Locke's *Two Treatises of Government* is an outstanding example of literature written to justify the rights of man against absolutism. This growth of abstract theory in the 17th century, this increasing tendency to construct systems and discuss politics in terms of principles, marks the emergence of the ideological style. In political conversation generally it was accompanied by a growing use of concepts such as right and liberty—ideals in terms of which actual policies were judged.

Hegel and Marx. Although the word *ideology* in the sense derived from Destutt de Tracy's understanding has passed into modern usage, it is important to notice the particular sense that *ideology* is given in Hegelian and Marxist philosophy, where it is used in a pejorative way. *Ideology* there becomes a word for what these philosophers also call "false consciousness." G.W.F. Hegel argued that people were instruments of history; they enacted roles that were assigned to them by forces they did not understand; the meaning of history was hidden from them. Only the philosopher could expect to understand things as they were. This Hegelian enterprise of interpreting reality and reconciling the world to itself was condemned by certain critics as an attempt to provide an ideology of the status quo, in that if individuals were indeed mere ciphers whose actions were determined by external forces, then there was little point in trying to change or improve political and other circumstances. This is a criticism Karl Marx took up, and it is the argument he developed in *The German Ideology* and other earlier writings. Ideology in this sense is a set of beliefs with which people deceive themselves; it is theory that expresses what they are led to think, as opposed to that which is true; it is false consciousness.

Marx, however, was not consistent in his use of the word *ideology,* for he did not always use the term pejoratively, and some of his references to it clearly imply the possibility of an ideology being true. Twentieth-century Marxists, who have frequently discarded the pejorative sense of ideology altogether, have been content to speak of Marxism as being itself an ideology. In certain Communist countries "ideological institutes" have been established, and party philosophers are commonly spoken of as party ideologists. Marxism is an excellent example, a paradigm, of an ideology.

The sociology of knowledge. The use of the word *ideology* in the pejorative sense of false consciousness is found not only in the writings of Marx himself but in those of other exponents of what has come to be known as the sociology of knowledge, including the German sociologists Max Weber and Karl Mannheim, and numerous lesser figures. Few such writers are wholly consistent in their use of the term, but what is characteristic of their approach is their method of regarding idea systems as the outcome or expression of certain interests. In calling such idea systems ideologies, they are treating them as things whose true nature is concealed; they consider the task of sociological research to be the unveiling of what Mannheim called the "life conditions which produce ideologies."

From this perspective, the economic science of Adam Smith, for example, is not to be understood as an independent intellectual construction or to be judged in terms of its truth, consistency, or clarity; rather, it is to be seen as the expression of bourgeois interests, as part of the ideology of capitalism.

Freudian support

The sociology of knowledge in its more recent formulations has sought support in Freudian psychology (notably in borrowing from Freud the concepts of the unconscious and of rationalization), in order to suggest that ideologies are the unconscious rationalizations of class interests. This refinement has enabled sociologists of knowledge to rid their theory of the disagreeable and unscientific element of bald accusation; they no longer have to brand Adam Smith as a deliberate champion of the bourgeois ethos but can see him now as simply the unconscious spokesman of capitalism. At the same time, these sociologists of knowledge have argued that Freudian psychology is itself no less a form of ideology than is Adam Smith's economics, for Freud's method of psychoanalysis is essentially a technique for adjusting rebellious minds to the demands and constraints of bourgeois society.

Critics of the sociology of knowledge have argued that if all philosophy is ideology, then the sociology of knowledge must itself be an ideology like any other idea system and equally devoid of independent validity; that if all seeming truth is veiled rationalization of interest, then the sociology of knowledge cannot be true. It has been suggested that although Weber and Mannheim inspired most of the work that has been done by sociologists of knowledge their own writings may perhaps be exempted from this criticism, if only on the ground that neither of them put forward a consistent or unambiguous theory of ideology. Both used the word *ideology* in different ways at different times. Weber was in part concerned to reverse Marx's theory that all idea systems are products of economic structures, by demonstrating conversely that some economic structures are the product of idea systems (that Protestantism, for example, generated capitalism and not capitalism Protestantism). Mannheim, on the other hand, tried to restore in a more elaborate form Marx's suggestion that ideologies are the product of the social structure. But Mannheim's analysis may have been obscured by his proposal that the word *ideology* should be reserved for idea systems that are more or less conservative, and the word *utopia* for idea systems of a more revolutionary or millenarian nature. Mannheim did not, however, remain faithful to this stipulative definition, even in his book entitled *Ideology and Utopia.*

On the other hand, Mannheim was well aware of the implication of the doctrine that all idea systems have a class basis and a class bias. As a way out of the dilemma he envisaged the possibility of a classless class of intellectuals, a "socially unattached intelligentsia," as he put it, capable of thinking independently by virtue of its independence from any class interest or affiliation. Such a detached group might hope to acquire knowledge that was *not* ideology.

This vision of a small elite of superior minds rising above the myths of ordinary society seemed to some readers to put Mannheim closer to Plato than to Marx and to cast new doubts on the claim of the sociology of knowledge to be a science.

THE POLITICAL CONTEXT

Ideology, rationalism, and romanticism. If some theorists emphasize the kinship between ideology and various forms of religious enthusiasm, others stress the connection between ideology and what they call rationalism, or the attempt to understand politics in terms of abstract ideas rather than of lived experience. Like Napoleon, who held that ideology is par excellence the work of intellectuals, some theorists are suspicious of those who think they know about politics because they have read many books; they believe that politics can be learned only by an apprenticeship to politics itself.

Such people are not unsympathetic to political theories, such as Locke's, but they argue that their value resides in the facts that are derived from experience. Michael Oakeshott in England has described Locke's theory of political liberty as an "abridgment" of the Englishman's traditional understanding of liberty, and has suggested that once such a conception is uprooted from the tradition that has given it meaning it becomes a rationalistic doctrine or metaphysical abstraction, like those liberties contained in the Declaration of the Rights of Man, which were so much talked about after the French Revolution but rarely actually enjoyed, in France or elsewhere.

Whereas Oakeshott has seen ideology as a form of rationalism, Edward Shils, a U.S. political scientist, has seen it more as a product of, among other things, romanticism with an extremist character. His argument is that romanticism has fed into and swelled the seas of ideological politics by its cult of the ideal and by its scorn for the actual, especially its scorn for what is mediated by calculation and compromise. Since civil politics demands both compromise and contrivance and calls for a prudent self-restraint and responsible caution, he suggests that civil politics is bound to be repugnant to romanticism. Hence Shils concludes that the romantic spirit is naturally driven toward ideological politics.

Existential critique of ideology

Ideology and terror. The "total" character of ideology, its extremism and violence, have been analyzed by other critics, among whom the French philosopher-writer Albert Camus and the Austrian-born British philosopher Sir Karl Popper merit particular attention. Beginning as an Existentialist who subscribed to the view that "the universe is absurd," Camus passed to a personal affirmation of justice and human decency as compelling values to be realized in conduct. An Algerian by birth, Camus also appealed to what he believed to be the "Mediterranean" tradition of moderation and human warmth and joy in living as opposed to the "northern" Germanic tradition of fanatical, puritan devotion to metaphysical abstractions. In his book *The Rebel* (*L'Homme révolté*), he argued that the true rebel is not the man who conforms to the orthodoxy of some revolutionary ideology, but a man who could say "no" to injustice. He suggested that the true rebel would prefer the politics of reform, such as that of modern trade-union socialism, to the totalitarian politics of Marxism or similar movements. The systematic violence of ideology—the *crimes de logique* that were committed in its name—appeared to Camus to be wholly unjustifiable. Hating cruelty, he believed that the rise of ideology in the modern world had added enormously to human suffering. Though he was willing to admit that the ultimate aim of most ideologies was to diminish human suffering, he argued that good ends did not authorize the use of evil means.

A somewhat similar plea for what he called "piecemeal social engineering" was put forward by Popper, who argued that ideology rests on a logical mistake: namely the notion that history can be transformed into science. In *The Logic of Scientific Discovery* (*Logik der Forschung*), Popper suggested that the true method of science was not one of observation, hypothesis, and confirmation but one of conjecture and experiment, in which the concept of falsification played a crucial role. By this concept he meant that in science there is a continuing process of trial and error; conjectures are put to the test of experiment, and those that are not falsified are provisionally accepted; thus there is no definitive knowledge but only provisional knowledge that is constantly being corrected. Popper saw in the enterprise of ideology an attempt to find certainty in history and to produce predictions on the model of what were supposed to be scientific predictions. Ideologists, he argued, because they have a false notion of what science is, can produce only prophecies, which are quite distinct from scientific predictions and which have no scientific validity whatever. Though Popper was well disposed toward the idea of a "scientific" approach to politics and ethics, he suggested that a full awareness of the importance of trial and error in science would prompt one to look for similar forms of "negative judgment" elsewhere.

By no means are all ideologists explicit champions of violence, but it is characteristic of ideology both to exalt action and to regard action in terms of a military analogy. Some observers have pointed out that one has only to consider the prose style of the founders of most ideologies to be struck by the military and warlike language that they habitually use, including words like *struggle, resist, march, victory,* and *overcome;* the literature of ideology is replete with martial expressions. In such a view, commitment to an ideology becomes a form of enlistment so that to become the adherent of an ideology is to become a combatant or partisan.

In the years that followed World War II, a number of ideological writers went beyond the mere use of military language and made frank avowals of their desire for violence—not that it was a new thing to praise violence. The French political philosopher Georges Sorel, for example, had done so before World War I in his book *Reflections on Violence.* Sorel was usually regarded as being more a Fascist than a Socialist. He also used the word *violence* in his own special way; by violence Sorel meant passion, not the throwing of bombs and the burning of buildings.

Violence found eloquent champions in several black militant writers of the 1960s, notably the Martinican theorist Frantz Fanon. Moreover, several of the French philosopher Jean-Paul Sartre's dramatic writings turn on the theme that "dirty hands" are necessary in politics and that a man with so-called bourgeois inhibitions about bloodshed cannot usefully serve a revolutionary cause. Sartre's attachment to the ideal of revolution tended to increase as he grew older, and in some of his later writings he suggested that violence might even be a good thing in itself.

Sartrian ideology

In considering Sartre's views on the subject of ideology it must be noted that Sartre sometimes used the word *ideology* in a sense peculiarly his own. In an early section of his *Search for a Method* (*Critique de la raison dialectique*), Sartre drew a distinction between philosophies and ideologies in which he reserved the term *philosophy* for those major systems of thought, such as the Rationalism of Descartes or the Idealism of Hegel, which dominate men's minds at a certain moment in history. He defined an ideology as a minor system of ideas, living on the margin of the genuine philosophy and exploiting the domain of the greater system. What Sartre proposed in this work was a revitalization and modernization of the "major philosophy" of Marxism through the integration of elements drawn from the "ideology," or minor system, of Existentialism. What emerged from the book was a theory in which the Existentialist elements are more conspicuous than the Marxist.

Ideology and pragmatism. A distinction is often drawn between the ideological and the pragmatic approach to politics, the latter being understood as the approach that treats particular issues and problems purely on their merits and does not attempt to apply doctrinal, preconceived remedies. Theorists have debated whether or not politics has become less ideological and whether a pragmatic approach can be shown to be better than an ideological one.

On the first question, there seemed to be good reason for thinking that after the death of Stalin and the repudiation of Stalinism by the Communist Party, the Soviet Union, at least, was becoming more interested in the "pragmatic"

concerns of national security and the balance of power and less interested in the ideological aim of fostering universal Communism. This in turn seemed to many to have resulted—in both the United States and the Soviet Union—in a shift toward a pragmatic policy of coexistence and a peaceful division of spheres of influence. There were indications in many countries that the old antagonisms between capitalist and socialist ideologies were giving way to a search for techniques for making a mixed economy work more effectively for the good of all.

But while many observers believed that there was much evidence of a decline of ideology in the latter 1950s, others believed that there were equally manifest signs in the following decade of a revival of ideology, if not within the major political parties, then at least among the public generally. Throughout the world various left-wing movements emerged to challenge the whole ethos on which pragmatic politics was based. Not all these ideologies were coherent, and none possessed the elaborate intellectual structure of the 19th-century ideologies; but together they served to demonstrate that the end of ideology was not yet at hand.

Pragmatism

As suggested earlier, certain controversies about ideology have to some extent been rooted in the ambiguity of the word itself, and this is perhaps especially relevant to the confrontation between ideology and pragmatism, since the word *pragmatism* raises problems no less intractable than those involved in connection with the word *ideology*. In the senses outlined at the beginning of this article, ideology is manifestly not the only alternative to pragmatism in politics, and to reject ideology would not necessarily be to adopt pragmatism. Ordinary language does not yet yield as many words as political science needs to clarify the question, and it becomes necessary to introduce such expressions as belief system, or to name the relevant distinctions, to further the analysis.

Almost any approach to politics constitutes a belief system of one kind or another. Some such belief systems are more structured, more ordered, and generally systematic than others. Though an ideology is a type of belief system, not all belief systems are ideologies. One man's belief system may consist of a congeries of ill-assorted prejudices and inarticulate assumptions. Another's may be the result of deep reflection and careful study. It is sometimes felt to be convenient to speak of a belief system of this latter type as a philosophy or, better, to distinguish it from philosophy in the technical or academic sense, as a *Weltanschauung* (literally, a "view of the world").

The confrontation between ideology and pragmatism may be more instructive if it is translated into a distinction between the ideological and the pragmatic, taking these two adjectives as extremes on a sliding scale. From this perspective, it becomes possible to speak of differences of degree, to speak of an approach to politics as being more or less ideological, more or less pragmatic. At the same time it becomes possible to speak of a belief system such as liberalism as lending itself to a variety of forms, tending at the one extreme toward the ideological, and at the other toward the pragmatic.

THE CONTEXT OF INTERNATIONAL RELATIONS

Dominance of contemporary international relations by ideology

It has been said that ideology has transformed international relationships in the 20th century—in appearance at least. Earlier centuries experienced dynastic wars, national, civil, and imperial wars, and diplomacy designed to further national security or national expansion or to promote mutual advantages and general peace. Such factors, indeed, appeared to govern international relations until recent times. International relations today are seemingly dominated more often than not by the exigencies of "-isms": wars are fought, alliances are made, and treaties are signed because of ideological considerations. The balance of power in the contemporary world is a balance weighted by ideological commitment. "The Communist bloc" confronts "the Free peoples," and in the "Third World" emergent nations cultivate a nationalist, anticolonialist ideology in their search for identity and their efforts to achieve modernity.

But this is not to assert that ideological wars, or ideological diplomacy, are entirely new. What has become the most conspicuous element in contemporary international relations—so conspicuous that other elements are often entirely ignored—was present, to a lesser degree, in earlier international relations. It is necessary here to distinguish between the actual events of history and the interpretations that are put on history, for some events lend themselves more readily than others to an ideological interpretation. The ideological perspective has become increasingly significant as the general public has come to play a role in considering questions of war and peace. When questions of defense and diplomacy were settled by kings and their ministers and wars were fought by professional soldiers and sailors, the public was not expected to have any opinion about international relations, and in such a situation there was little place for ideology.

Ideology in the World Wars. In the course of World War I, however, a new element appeared to have been introduced. The war was seen by those who experienced it as being in its early stages a national war of the traditional kind, and as such it was not at first expected to assume any profoundly disturbing form. Each combatant people viewed itself as fighting for king and country in a just war. But by 1916 the Allies were being urged to think of their endeavour as a war "to make the world safe for democracy," and the Germans, on their side, were correspondingly encouraged to visualize the war as a struggle of "culture" against "barbarism." On both sides, the casualties were far more terrible than anyone had foreseen, and the need to sustain the will to war by an appeal to ideology was plainly felt by all the nations involved. Whether such "war aims" were really the main objectives of the governments concerned is another question; what is important is that, as the need was increasingly felt for a justification of war, the justification took an ideological form. Whether or not World War I changed its real nature between 1914 and 1918, the prevailing conception of it underwent significant alteration. This became more marked after the Russian Revolution of 1917, when the Bolsheviks submitted to harsh German peace terms for reasons that were not only practical but ideological—namely, the preservation and promotion of Communism. Pres. Woodrow Wilson took the United States into the war on the Allied side with an alternative ideological vision—that of ensuring permanent peace through the League of Nations and of establishing democratic governments in all the conquered countries.

Ideology and the will to fight

The rise of Communism clearly marked a corresponding increase in the role of ideology in international relations. Fascism helped to speed the process. The Spanish Civil War of the 1930s was an almost clear-cut confrontation between the ideologies of left and right (not entirely clear-cut because of the ambiguous relationship between Communism and anarchism).

The precise extent of ideological commitment in World War II is a matter of some controversy. At one level, the 1939 war is seen as a continuation of the war of 1914. Two of the leading protagonists—Great Britain and the United States—agreed more in their anti-ideological stance and their hostility to Nazism than in promoting an alternative ideology. Pres. Franklin D. Roosevelt, suspicious of British and French imperialism and eager to cultivate a progressive ideological outlook, was critical of Prime Minister Winston Churchill's politics, hostile toward Charles de Gaulle's, but surprisingly tolerant of Joseph Stalin's. The revival of Wilson's idealistic war aims in the Atlantic Charter provided a basis for a kind of general ideological union of the Allies. But such formulations proved to be of small significance compared to the profound ideological commitment of the Soviet Union to Communism, and that of the United States to an international position more ideologically anti-Communist than pro anything.

Ideology of the Cold War. What came to be called the Cold War in the 1950s must be understood, to a large extent, as an ideological confrontation, and, whereas Communism is manifestly an ideology, the "non-Communism," or even the "anti-Communism," of the West is negatively ideological. To oppose one ideology is not necessarily to subscribe to another, although there is a strong body of opinion in the West that feels that the free world needs a coherent ideology if it is to resist successfully an opposing ideology.

"Anti-" and "non-"

The connection between international wars and ideology can be better expressed in terms of a difference of degree rather than of kind: some wars are more ideological than others, although there is no clear boundary between an ideological and nonideological war. An analogy with the religious wars of the past is evident, and there is indeed some historical continuity between the two types of war. The Christian Crusades against the Turks and the wars between Catholics and Protestants in early modern Europe have much in common with the ideological conflicts of the contemporary period. Religious wars are often communal wars, as witness those between Hindus and Muslims in India; but an "ideological" element of a kind can be discovered in many religious wars, even those narrated in the Old Testament, in which the people of Israel are described as fighting for the cause of righteousness—fighting, in other words, for a universal abstraction as distinct from a local and practical aim. In the past this "ideological" element has in the main been subsidiary; what is characteristic of the modern period is that the ideological element has become increasingly dominant, first in the religious wars (and the related diplomacy) that followed the Reformation, and then in the political wars and diplomacy of recent times.

BIBLIOGRAPHY. A useful introduction is provided by MARTIN SELIGER, *Ideology and Politics* (1976), which works from a broad definition of the concept of ideology. JOHN PLAMENATZ, *Ideology* (1970), is a clear and uncomplicated study by a distinguished Oxford philosopher. JEAN BAECHLER, *Qu'est-ce que l'idéologie* (1976), is characteristically French in his approach and affords an equally lucid introduction to both the sociological and the historical aspects of the problem. Another book written at a fairly popular level is PATRICK CORBETT, *Ideologies* (1966).

Few of the works of the original French *idéologues* are available in modern editions and even fewer in English translations. However, RICHARD H. COX (ed.), *Ideology, Politics, and Political Theory* (1969), contains short excerpts in English translation from Destutt de Tracy and his contemporaries as well as from more recent works. DESTUTT DE TRACY's major work in the field is *Élémens d'idéologie,* 2nd ed. (1805–18); and the expository study by F.J. PICAVET, *Les Idéologues* (1891, reprinted 1975), remains a classic.

GEORGE LICHTHEIM, *The Concept of Ideology* (1967), contains a short but well-informed and sympathetic analysis of ideology as it figures in Hegelian and Marxist thought. Relevant elements of Hegel's philosophy may be found in his *Lectures on the Philosophy of History* (1956; originally published in German, 1837). Valuable commentaries are provided by CHARLES TAYLOR, *Hegel* (1975); JEAN HYPPOLITE, *Studies on Marx and Hegel* (1969; originally published in French, 1955); and ALEXANDRE KOJÈVE, *Introduction to the Reading of Hegel* (1969, reissued 1980; originally published in French, 1947). In Marxist philosophy, KARL MARX and FRIEDRICH ENGELS, *The German Ideology* (1965; originally published in German, 1932), is the fundamental text. Recent treatments of ideology in the Marxist tradition include JOE MCCARNEY, *The Real World of Ideology* (1980); JORGE LARRAIN, *The Concept of Ideology* (1979); COLIN SUMNER, *Reading Ideologies: An Investigation into the Marxist Theory of Ideology and Law* (1979); and ALVIN W. GOULDNER, *The Dialectic of Ideology and Technology* (1976, reissued 1982).

Writers who have attempted to formulate a neo-Marxist theory of ideology, drawing in part on Hegelian philosophy, include KARL MANNHEIM, *Ideology and Utopia* (1936, reissued 1970; originally published in German, 1929); JÜRGEN HABERMAS, *Towards a Rational Society* (1971); and HERBERT MARCUSE, *One-Dimensional Man* (1964). JEAN-PAUL SARTRE, *Critique of Dialectical Reason* (1976; originally published in French, 1960), constructs a theory of ideology as a "marginal system of ideas" that is consciously designed as an alternative to Marxist theory.

Interpretations of ideology that are directly opposed to Marxist theory include LEWIS S. FEUER, *Ideology and the Ideologists* (1975); MARTIN SELIGER, *The Marxist Conception of Ideology* (1977); DAVID MANNING, *The Form of Ideology* (1980); and JAMES R. FLYNN, *Humanism and Ideology* (1973).

Historical studies that take a relatively extensive view of the impact of ideology as a revolutionary force in the modern world include JAMES H. BILLINGTON, *Fire in the Minds of Men* (1980); MELVIN J. LASKY, *Utopia and Revolution* (1976); and JEANNE HERSCH, *Idéologies et réalité* (1956). F.M. WATKINS, *The Age of Ideology: Political Thought 1750 to the Present,* 2nd ed. (1979); and HANS KOHN, *Political Ideologies of the Twentieth Century,* 3rd ed. rev. (1966), treat ideology as the dominant characteristic of modern political thinking. More polemical commentaries on the development of ideology include KARL POPPER, *The Poverty of Historicism* (1957); ALBERT CAMUS, *The Rebel* (1953; originally published in French, 1951); and JEAN-FRANÇOIS REVEL, *Pourquoi les philosophes?* (1957). A systematic critique of the whole notion of ideological politics may be found in the works of MICHAEL OAKESHOTT, notably *Rationalism in Politics* (1962, reissued 1981), *On Human Conduct* (1975, reissued 1983), and *On History* (1983).

RAYMOND ARON, *The Opium of the Intellectuals* (1957, reprinted 1977; originally published in French, 1955), points to a decline in ideological politics in the West; as does DANIEL BELL, *The End of Ideology* (1960). Less confident views are advanced in SIDNEY HOOK, *Pragmatism and the Tragic Sense of Life* (1974); and in DAVID APTER (ed.), *Ideology and Discontent* (1964).

Sociological aspects of ideology are explored in DONALD MACRAE, *Ideology and Society* (1961); NORMAN BIRNBAUM, *The Sociological Study of Ideology: 1940–1960* (1962); ERIC CARLTON, *Ideology and Social Order* (1977); and FRANÇOIS BOURRICAUD, *Le Bricolage idéologique* (1980). ARNE NAESS, *Democracy, Ideology and Objectivity* (1956), written from the perspective of political philosophy, was the first of a series of works that investigate the relationship between ideology and liberty. Among these may be mentioned DANTE GERMINO, *Beyond Ideology* (1967, reprinted 1976); Z.A. JORDAN, *Philosophy and Ideology* (1963); JUDITH N. SHKLAR (ed.), *Political Theory and Ideology* (1966); and MAURICE CRANSTON and PETER MAIR (eds.), *Ideology and Politics* (1980). K.R. MINOGUE, *Alien Power: The Pure Theory of Ideology* (1985), uses both a philosophical and a historical approach to provide a far-reaching survey of the subject. Among books that stay close to the main tradition of American political science, the following are notable: ROBERT LANE, *Political Ideology* (1962); WILLIAM E. CONNOLLY, *Political Science and Ideology* (1967); and ROBERT A. DAHL, *After the Revolution* (1970). ANDREW GYORGY and GEORGE D. BLACKWOOD, *Ideologies in World Affairs* (1967), analyzes the emergence of ideology as a decisive factor in international relations.

(M.C.)

Immunity

Immunity describes the ability of humans and other animals to resist or overcome infection by invading microbes or larger parasites. That the environment contains many potentially destructive microorganisms is obvious from the speed with which animals decay after death. It is impossible for an animal to avoid contact with microbes, many billions of which live harmlessly on the skin and in the gut, quite apart from those that are breathed in or that penetrate the skin whenever it is pricked or cut. What matters is that microbes should not grow once inside the animal's body. Humans and other vertebrates have two sorts of defenses against microbial infection. One consists of the so-called innate immune mechanisms, those that kill or inhibit a wide variety of microbes irrespective of whether these have challenged the body before. This kind of immunity is nonspecific—that is, its mechanisms can act against microbes that are not necessarily similar to one another. The other type of defense mechanisms, by contrast, provides specific, acquired immunity. Such immunity is specific in that its responses are tailored to act against a particular microbe or its products; it is acquired in that these tailor-made responses are enormously increased as a result of being stimulated by the prior presence of a given microbe or its products. Vaccination, or active immunization, confers protection by stimulating specific, acquired immunity.

Obviously any animal that has mechanisms to destroy invaders must be able to distinguish between the materials of its own body and those that originate outside it. Considering that all living creatures are made up of basically similar building blocks, the capacity of any creature to distinguish the molecules of which it is composed—*i.e.*, "self"—from practically all others—*i.e.*, "not-self"—is remarkable. This capacity is present in all living creatures to some degree, but among vertebrates it is especially a feature of the white blood cells called lymphocytes; much of this article focuses on these cells.

Despite its obvious benefits, specific, acquired immunity can become misdirected and do more harm than good. Allergies arise when the mechanisms of immunity become exaggerated or when they are directed against otherwise harmless, nonliving materials absorbed from outside, such as pollen, dust, or food. Autoimmune disorders occur when immune responses become directed against the body's own constituents.

For the most part, this article centres on the human immune system. The discussion is divided into the following sections:

Nonspecific, innate immunity 773
- Barriers against infection 773
- Protective factors 773
- Scavenger and natural killer cells 774
  - Scavenger cells
  - Natural killer (NK) cells
- Acute-phase reactions 774

Specific, acquired immunity 775
- The nature of lymphocytes 775
  - General characteristics
  - Ability to recognize foreign molecules
  - Diversity of lymphocytes
- Antibodies and B-cell antigen receptors 776
  - Basic structure
  - Classes of immunoglobulins
- T-cell antigen receptors 778
- Life cycle of T and B lymphocytes 778
- Activation of lymphocytes 778
  - Activation of T cells
  - Activation of B cells
- Regulation of immune responses 779
- Antibody-mediated immune mechanisms 779
  - Protective attachment to antigens
  - Activation of the complement system
  - Activation of killer cells
  - Other antibody-mediated mechanisms
  - Transfer of antibodies from mother to offspring
- Cell-mediated immune mechanisms 781
- Immunity against cancer 781
- Prophylactic immunization 782
  - Passive immunization
  - Active immunization
- Production of monoclonal antibodies 782

Disorders involving the immune system 783
- Immune deficiencies 783
- Allergies 784
  - Type I hypersensitivity
  - Type II hypersensitivity
  - Type III hypersensitivity
  - Type IV hypersensitivity
- Autoimmune disorders 786
  - Basic processes that underlie autoimmunity
  - Examples of autoimmune disorders
  - Treatment
- Cancers of the lymphocytes 788

Evolution of the immune system 788
- Development of immunity in major animal groups 788
  - Immune capacity among invertebrates
  - Immune capacity among vertebrates
- Genetic origins of the immune system 789

Bibliography 789

Nonspecific, innate immunity

BARRIERS AGAINST INFECTION

The skin and the linings of the respiratory and gastrointestinal tracts provide the first line of defense against invasion by microbes or parasites. Human skin has a tough outer layer of cells that produce keratin, a protein similar to that of which hair is composed. This layer of cells, which is constantly renewed from below, serves as a mechanical barrier to infection. In addition, glands in the skin secrete oily substances that include fatty acids, such as oleic acid, that can kill some bacteria; skin glands also secrete lysozyme, an enzyme (also present in tears) that can break down the outer wall of certain bacteria.

Like the outer layer of the skin, but much softer, the linings of the respiratory and gastrointestinal tracts provide a mechanical barrier of cells that are constantly being renewed. The lining of the respiratory tract has cells that secrete mucus (phlegm), which traps small particles. Other cells in the wall of the respiratory tract have small hairlike projections, called cilia, which steadily beat in a sweeping movement that propels the mucus and any trapped particles up and out of the throat and nose. As will be clear from later sections of this article, protective antibodies, which are products of specific immunity, are also present in the mucus. The lining of the gastrointestinal tract also contains cells that secrete mucus, which, in addition to aiding the passage of food, traps potentially harmful particles. This mucus also contains protective antibodies; furthermore, the stomach lining secretes hydrochloric acid that is strong enough to kill many microbes.

PROTECTIVE FACTORS

As stated above, some microbes do penetrate the body's protective barriers and enter the internal tissues. There they encounter a variety of chemical substances that may prevent their growth. Some of the chemicals are ones involved in normal body processes. For example, the blood

and tissue fluids contain chemicals that inhibit the potentially damaging digestive enzymes released from the body cells that have died in the natural course of events, but these chemicals can also inhibit similar enzymes produced by bacteria. Another substance that performs normal processes but also provides protection against microbes is the blood protein transferrin. This protein binds free molecules of iron firmly and transports the metal from the gut, where it is absorbed, to those cells that use it (*e.g.,* to make the red-blood-cell pigment hemoglobin). Bacteria need free iron to grow, and transferrin can render it unavailable to them. Probably more important than the "incidental" protective chemicals just described is the mixture of proteins termed complement, which is present in blood and tissue fluids and is discussed more fully later in this article. Many microbes can activate complement in ways that do not involve the special means used in specific immunity. If the microbes lack protective coats, complement will destroy them. Other microbes, including those that are potentially harmful (pathogenic), can evade these mechanisms. Protection against these invaders depends upon the activity of scavenger cells, together with the mechanisms to be described below under specific immunity.

nterferons

Yet another protective factor is a family of proteins collectively known as interferons; these inhibit the replication of many, but not all, viruses. Interferons are made normally by cells that are already infected by viruses or that have taken in foreign nucleic acid (nucleic acid being the essential constituent of viruses, which enables them to replicate within infected cells). When first discovered, interferon was thought to be a single substance, but scientists have learned that different cells produce different kinds. Alpha interferons are produced by white blood cells other than lymphocytes, beta interferon by fibroblasts, and gamma interferon by lymphocytes. All interferons inhibit viral replication by interfering with the transcription of viral nucleic acid, but they also exert other inhibitory effects by regulating the extent to which lymphocytes and other cells express some important constituents on their surface membranes and by stimulating the natural killer (NK) cells described below.

SCAVENGER AND NATURAL KILLER CELLS

Scavenger cells. All higher animals, and many lower ones, contain scavenger cells capable of the phagocytosis (ingestion) and destruction of foreign particles. Most vertebrates, including all birds and mammals, possess two main kinds of scavenger cells. Their importance was first recognized by the Russian biologist Élie Metchnikoff in 1884. He named them microphages and macrophages, after Greek words meaning "little eaters" and "big eaters."

Microphages are nowadays usually referred to either as polymorphonuclear leukocytes, because their nucleus typically has several lobes, or as granulocytes, because they contain large numbers of obvious granules. Some of these granules are packets containing a mixture of powerful digestive enzymes. Others contain bactericidal (*i.e.,* bacteria-killing) proteins that act by generating hypochlorite or hydrogen peroxide when particles have been ingested.

Types and functions of granulocytes

There are three varieties of granulocytes, distinguished by the way in which their granules are stained by dyes. The differences in staining characteristics reflect differences in the chemical makeup and function of the granules. The most common granulocytes are the neutrophils, which serve as general-purpose scavengers. Less common are the eosinophils, which are particularly effective at damaging the cells that make up the cuticle (body wall) of larger parasites. Fewer still are the basophils, which release heparin (a complex sugar that prevents blood coagulation) as well as histamine and other substances that play an important role in certain allergic reactions described later in this article.

Granulocytes are continuously produced from stem (*i.e.,* precursor) cells in the bone marrow, from which they enter the bloodstream, circulate for a few days, and then die. Granulocytes are mobile and are attracted toward foreign materials, including bacteria, which they phagocytize and in many cases digest. Bacteria deposited beneath the skin, for example, are attacked within a few minutes by neutrophils that move to the invasion site from nearby blood vessels. Various agents attract the different kinds of granulocytes. Some of these attracting agents are produced by the microbes themselves; others are released by damaged tissues. Still others result from the interaction of the microbes with proteins in the blood plasma—a particularly powerful stimulus is the activation of complement. Receptors, structures found on the surface of granulocytes, determine how these cells recognize foreign materials. The most important receptors are those that recognize and bind to antibody molecules, which have coated an invader, and those that bind to the activated form of complement. Although granulocytes make up a critical component of nonspecific immunity, some microbes produce toxins that poison granulocytes and thus prevent ingestion; other microbes are indigestible and thus are not killed. By themselves, then, granulocytes are of limited effectiveness, and they require reinforcement by the mechanisms of specific immunity.

Origin and function of macrophages

Macrophages, the other main type of phagocytic cell, are derived from cells called monocytes. Like granulocytes, monocytes are produced by stem cells in the bone marrow and circulate through the blood, though in lesser numbers. But unlike granulocytes, monocytes then take on a maturer form—that of macrophages—and settle in many tissues, especially the lymphoid tissues (spleen and lymph nodes) and the liver, which serve as filters for trapping microbes and other foreign particles that arrive through the blood or the lymph. Macrophages live longer than granulocytes, and, although effective as scavengers, they serve a longer term and basically different function. Compared to granulocytes, macrophages move relatively sluggishly and are attracted by different stimuli; they usually arrive at sites of invasion later than granulocytes. Macrophages recognize foreign particles by mechanisms basically similar to those of granulocytes and ingest them effectively. Ingested microbes are killed and digested by much the same means as employed by granulocytes, but the process is slower and digestion is not as complete. This aspect will be shown later in this article to be of great importance for the role of macrophages in stimulating specific immune responses—something in which granulocytes play no part.

Natural killer (NK) cells. Natural killer cells were first recognized in 1975, when researchers observed cells in the blood and lymphoid tissues that, although they were neither the scavengers described above nor ordinary lymphocytes, could kill some tumour cells and cells infected with viruses cultured in vitro. (In vitro, literally "in glass," is a Latin term used for studies conducted outside the body, as opposed to in vivo studies, which are conducted in living animals.) NK cells look like lymphocytes but contain a few granules resembling those of granulocytes. They can apparently recognize some feature of dividing cells, by a mechanism that does not depend upon specific immunity; they then bind to these dividing cells and insert their granules through the outer membrane and into the cytoplasm. This causes the dividing cells to leak and die. It is not certain whether NK cells belong to a distinct lineage or are a special form of lymphocyte, but there is no doubt that they are stimulated by gamma interferon. Most immunologists feel that NK cells play an important part in checking the growth of tumour cells and cells infected with some viruses. It is likely, however, that their main biological role is to regulate the growth of stem cells in the bone marrow and elsewhere.

ACUTE-PHASE REACTIONS

Infections are commonly accompanied by a rise in body temperature, or fever. This may sometimes be protective in itself, since some microbes cannot tolerate an increase of 2°–3° C (3°–5° F) above the normal body temperature. But fever composes part of a more general effect known as the acute-phase reaction, which stems from the production of a protein termed interleukin 1 (IL-1). IL-1, which is made by infected or otherwise activated macrophages, stimulates liver cells to secrete increased amounts of several different proteins into the bloodstream. These proteins bind to products released by dying tissue cells and assist in their harmless disposal; the proteins also help in repair

Interleukin 1

processes. The acute-phase reaction occurs in all warm-blooded animals that have been examined, and it can be triggered by tissue injury as well as by infections. Although not responsible for the actual elimination of infections, the acute-phase reaction is evidently an important part of the body's response once an infection has progressed so far as to make a person feel ill.

Specific, acquired immunity

It has been known for centuries that persons who have contracted certain diseases and survived generally do not catch these illnesses again. The Greek historian Thucydides recorded that when the plague was raging in Athens during the 5th century BC, the sick and dying would have received no nursing at all had it not been for the devotion of those who had already had the disease and recovered from it, since it was known that no one ever caught plague a second time. The same applies, with rare exceptions, to many other clear-cut diseases, such as smallpox, chicken pox, measles, and mumps. Yet having had measles does not prevent a child from contracting chicken pox, or vice versa. The protection acquired by experiencing one of these infections is specific for that infection; in other words, it is due to specific, acquired immunity. There are other infectious conditions, such as the common cold, influenza, pneumonia, and diarrheal diseases, that can be caught again and again; these seem to contradict the notion of specific immunity. But the reason why such illnesses can recur is that many different infectious agents produce similar symptoms (and thus the same disease). For example, more than 100 viruses can cause the cluster of symptoms known as the common cold. Consequently, even though an infection with any particular agent does protect against reinfection by that pathogen, people are still susceptible to the other agents that they have not yet encountered.

This section describes the various ways in which specific immunity operates. Although pioneer studies were begun in the late 19th century, most of the knowledge of specific immunity has been gained since the 1960s, and new insights are continually being obtained.

THE NATURE OF LYMPHOCYTES

General characteristics. As previously mentioned, lymphocytes are the cells responsible for the body's ability to distinguish and react to an almost infinite number of different foreign substances, including those of which microbes are composed. Lymphocytes are mainly a dormant population, awaiting the appropriate signals to be stirred to action. These inactive lymphocytes are small, round cells filled largely by a nucleus. Although they have only a small amount of cytoplasm compared with other cells, each lymphocyte has sufficient cytoplasmic organelles (mitochondria, endoplasmic reticulum, Golgi apparatus) to keep the cell alive. Lymphocytes move only sluggishly on their own, but they can travel around the body, carried along in the blood or the lymph. At any one time an adult human possesses about 2×10^{12} lymphocytes, about 1 percent of which are in the bloodstream. The great majority are concentrated in various tissues, particularly the bone marrow, spleen, thymus, lymph nodes, tonsils, and lining of the intestines. Organs or tissues containing such concentrations of lymphocytes are termed lymphoid. The lymphocytes in them are free to move but are not simply lying loose; rather, they are confined within a delicate network of connective tissue that channels the lymphocytes so that they come into contact with other cells, especially macrophages, that line the meshes of the network. This ensures that the lymphocytes—which, as discussed below, are of different sorts—come into contact with each other and with foreign materials trapped by the macrophages in an ordered manner.

Functions of lymphoid tissues

Lymphocytes originate, in postnatal life, from stem cells in the bone marrow; these stem cells divide continuously, releasing immature lymphocytes into the bloodstream. Some of these travel to the thymus, where they multiply and differentiate (*i.e.,* acquire special properties and functions) into T lymphocytes. The term T lymphocyte (or T cell) stands for thymus-derived lymphocyte (or cell), referring to the fact that these cells depend upon the maturation process that takes place in the thymus. Once they have left the thymus, T cells join the bloodstream and circulate to and within the rest of the lymphoid organs, where they can multiply further in response to appropriate stimulation. About half of all lymphocytes are T cells.

The other lymphocytes do not pass through the thymus; instead, they differentiate within the bone marrow itself and then go directly to the lymphoid organs. They are termed B lymphocytes, or B cells, and they, like T cells, can mature and multiply further in the lymphoid organs when suitably stimulated. Although it is appropriate to refer to them as B cells in humans and other mammals, because they are bone-marrow derived, the term B cell was originally used to describe the functionally analogous lymphocytes of birds, cells that differentiate not in the bone marrow but in the bursa of Fabricius, a special tissue attached to the hindgut.

Different functions of T cells and B cells

B cells and T cells differ in function. B cells secrete antibodies, proteins that attack foreign molecules. Since antibodies circulate through the humours (*i.e.,* the body fluids—blood and lymph), the protection afforded by B cells is called humoral immunity. T cells, by contrast, do not produce antibodies. Instead, the T lymphocytes themselves directly attack specific invaders and are thus said to provide cell-mediated immunity. The distinction between these two types of specific, acquired immunity is not so clean as might be inferred from the above, however, for T cells also play a major role in regulating the function of B cells. In many cases, moreover, an immune response involves both humoral and cell-mediated assaults upon the foreign substance.

Ability to recognize foreign molecules. The feature that distinguishes lymphocytes from other cells is their capacity to recognize foreign molecules by virtue of the three-dimensional patterns formed by the atoms of the foreign agent. To recognize a molecule, in the sense used throughout this article, a cell must have receptor molecules on its own surface. The shape of the receptors is complementary to a pattern on the surface of the molecule that is to be recognized. The great pioneer of immunology, Paul Ehrlich, used the analogy of a key fitting into a lock, and this provides a reasonably accurate and nontechnical way of imagining it—especially if the key and the lock are both regarded as made of rubber and thus somewhat flexible.

Antigens and immune response

At this stage it is necessary to introduce some terms used by immunologists. An antigen is any material, usually of a complex nature, that can bind specifically to receptor molecules made by lymphocytes. If the antigen also stimulates the lymphocytes to which it binds—*i.e.,* it evokes an immune response—it is said to be an immunogen. Complex antigens have a variety of different three-dimensional patterns on different parts of their surface, each of which can be recognized separately by different lymphocytes. Each of these recognizably foreign molecular patterns constitutes an antigenic determinant, or epitope. Complex antigens thus present an "antigenic mosaic" and can evoke responses from a variety of specific lymphocytes. Some antigenic determinants are more effective than others in stimulating a response, presumably because a larger number of responsive lymphocytes is present. When two different antigens have an antigenic determinant in common, some of the responses made against one antigen are able to react against the other; such antigens are known as cross-reacting antigens. A simple chemical group that can combine with a lymphocyte receptor but does not induce an immune response is called a hapten (Greek *haptein,* "to grasp"). Haptens can become immunogenic if they are joined to a suitable complex molecule, and immunologists often utilize haptens to study various immune reactions.

As will be described below, a B lymphocyte manufactures antibodies that correspond to its specific receptor molecule. It sheds these proteins in large amounts into the body fluids, where they can be detected by their ability to combine with antigens; in fact, antibodies were discovered in the blood long before it was known what cells made them. Collectively the antibodies are known as immunoglobulins. T lymphocytes secrete little or none of

their receptors into the blood, and there is no collective name for these.

Diversity of lymphocytes. The specific immune system (in other words, the sum total of all the lymphocytes) can recognize virtually any complex molecule that nature or science has devised. This very remarkable ability is achieved because among the population of lymphocytes, which are being lost and renewed throughout life at a rate of some 1 percent every day, each cell develops its own specific antigen receptor, and the antigen receptor of one lymphocyte is slightly different from that of most other lymphocytes. The total number of different receptors available on all of the lymphocytes cannot be measured, but most informed estimates place the total number at a minimum of at least 100,000,000. Each receptor is structurally organized so that it can combine with a particular antigenic determinant, so that among all the lymphocytes there are at least some that can recognize any given antigen. After a lymphocyte encounters an antigen that it can recognize, the cell is stimulated to multiply, and the population of lymphocytes bearing that particular receptor increases.

Genetic coding of antigen receptors

The antigen receptors on a lymphocyte are proteins, and, like all proteins, they are assembled from structural building blocks called amino acids. And, as with any protein produced by any cell, the exact sequence in which the amino acids of a particular antigen receptor are assembled is specified in the deoxyribonucleic acid (DNA) that makes up the cell's genes. Since the antigen receptors on different lymphocytes each have a different—even if only slightly different—sequence, this might imply that a person's genome (the sum total of all his genes, which are common to all cells in the body except the germ cells) must contain 100,000,000 genes coding for nothing but the possible forms of antigen receptors. This far exceeds the number of genes that all the DNA in a human cell could represent, and so it must be possible to code for all of the different receptor molecules with many fewer genes than would be expected. Recent work in molecular genetics has revealed how this is done, in a way that is peculiar to lymphocytes. (To learn more about the genetic control of protein synthesis in general, see GENETICS AND HEREDITY, THE PRINCIPLES OF.)

It turns out that the protein that makes up any specific antigen receptor is coded for not by a single gene but by different genes that specify separately the sequence of amino acids in four consecutive segments of the molecule. These genes come together to provide the code that ultimately specifies the whole molecule. For each segment of the molecule encoded separately there are up to 100 alternative genes, any of which may come together with any of the genes coding for the other segments to form the final code for the whole molecule. This rearrangement of alternative genes in the final code occurs largely, though not wholly, at random, so that an enormous number of combinations can result. The genetic rearrangement takes place at the stage when the lymphocytes generated from the stem cells first become functional, and so each lymphocyte begins its active life able to make antigen receptors composed of one particular protein out of all the possible variations. Consequently, within the whole population of lymphocytes there are nearly always some that can recognize any antigen. But this is not the only way in which the diversity of receptors is achieved. Superimposed on the mechanism outlined in simplified terms above is another process, called somatic mutation. This process involves mutation—small changes in the DNA that occur spontaneously during the process of cell division (when the DNA has to be copied into the daughter cells); it is called somatic because the mutations take place in body cells (Greek *soma*) rather than in germ-line cells. Although somatic mutations can develop as a chance event in any body cell, they occur regularly in the DNA that codes for antigen receptors in lymphocytes. Thus, when a lymphocyte is stimulated by an antigen to divide, new variants of its antigen receptor can be present on its descendant cells, and some of these variants may provide an even better fit for the antigen that was responsible for the original stimulation.

Somatic mutation and diversity of receptors

ANTIBODIES AND B-CELL ANTIGEN RECEPTORS

The antigen receptors on B lymphocytes are identical to the antibodies that these lymphocytes manufacture once stimulated, except that the receptor molecules have an extra tail that penetrates the cell membrane and anchors them to the cell surface. Thus, a description of the structure and properties of antibodies (immunoglobulins), which are well studied, will suffice for both.

Basic structure. Immunoglobulins (abbreviated Ig) all have the same basic molecular structure, consisting of four polypeptide chains (*i.e.*, chains of amino acids linked together by chemical bonds known as peptide bonds). Two of the chains, which are identical in any given immunoglobulin molecule, are heavy (H) chains; the other two are identical light (L) chains. The terms "heavy" and "light" simply mean larger and smaller, but they were used in early studies of immunoglobulin structure and have stuck. Each chain is manufactured separately and is coded for by different genes, but the four chains become joined in the final immunoglobulin molecule to form a flexible Y-shape as illustrated diagrammatically in Figure 1.

From B. Alberts *et al.*, *Molecular Biology of the Cell*; Garland Publishing, Inc., 1983

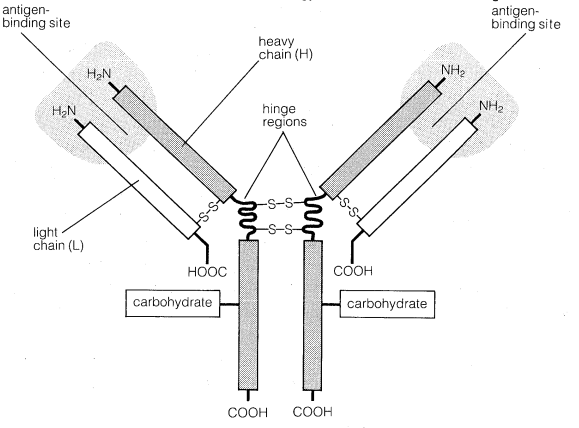


Figure 1: The basic four-chain unit of a human immunoglobulin molecule. This basic structure is composed of two identical light (L) chains and two identical heavy (H) chains, which are held together by disulfide (—S—S—) linkages to form a flexible Y-shape. The antigen-binding sites lie at the ends of each arm of the Y and are formed by a portion of both a light and a heavy chain.

Binding of antibody to antigen

At the end of each arm of the Y-shape, a portion of the H and L chains forms a region called the antigen-binding, or combining, site; each basic immunoglobulin molecule therefore has two identical antigen-binding sites. Their shape is such that part of the site can fit with a particular antigenic determinant. When two molecules fit and can approach very closely to one another, they become bound together by various weak forces. Although such forces are not as powerful as those of true chemical bonds (which bind atoms into molecules), they are nevertheless strong enough to hold the molecules together quite firmly. The two antigen-binding sites are found in those parts of the immunoglobulin molecule that differ between one antibody and another (as explained above) and are termed the variable (V) regions. The actual antigen specificity of the site depends upon the variable regions of both the H and L chains. In contrast to the variable regions, the other segments of the chains are alike from one immunoglobulin molecule to another. These segments are known as the constant (C) regions.

The variable and constant regions of both the L and H chains are structurally folded into functional units called domains. As shown in Figure 2, each light chain consists of one variable domain (V_L) and one constant domain (C_L). Each heavy chain has one variable domain (V_H) and three or four constant domains (C_H1, C_H2, C_H3, C_H4). Those domains that make up the "tail" of the basic Y-shaped molecule (in other words, all the H-chain constant domains except C_H1) are responsible for the special biological properties of immunoglobulins—except, of course, for the capacity to bind to a specific antigenic determinant.

Classes of immunoglobulins. The term "constant re-

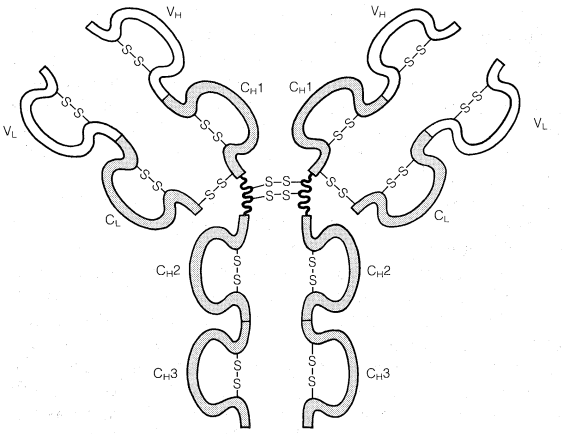


Figure 2: A human immunoglobulin molecule showing the domains present in the light (L) and heavy (H) chains. The constant (C) domains are shaded, while the variable (V) domains are not. Intrachain disulfide (—S—S—) linkages maintain the folding of the molecule into domains. The variable domains account for the specific antigen-binding properties of the immunoglobulin. The constant domains of the H chain confer other biological properties, such as the ability to bind to complement and to certain nonspecific receptors on cells.

From B. Alberts *et al.*, *Molecular Biology of the Cell*; Garland Publishing, Inc., 1983

gion" is a bit misleading in that these segments are not identical in all immunoglobulins. Rather, they are basically similar among broad groups. All immunoglobulins that have the same basic kinds of constant domains in their H chains are said to belong to the same class. There are five main classes—IgG, IgM, IgA, IgD, and IgE—each of which has special properties. Some of these main classes include a number of distinct subclasses. In addition, there are two basic kinds of L chains, either of which can be associated with any of the H-chain classes, thereby increasing still further the enormous diversity of immunoglobulins. These classes are illustrated diagrammatically in Figure 3.

Main classes of immunoglobulins

IgG is the most common class of immunoglobulins. IgG antibodies are present in the largest amounts in blood and tissue fluids and are made by B cells that have been previously stimulated by antigens. Each IgG molecule consists of the basic four-chain immunoglobulin structure and thus carries two identical antigen-binding sites. There are four subclasses of IgG, each with minor differences in their H chains but with distinct biological properties.

IgM is the earliest class of immunoglobulins made by B cells as they mature and is the form most commonly present as the antigen receptor on their surface. When IgM is secreted from the cells, five of the basic four-chain units become joined together to make a large molecule with 10 antigen-binding sites. This large antibody molecule is particularly effective at attaching to antigenic determinants present on the outer coats of bacteria.

IgA is a class of immunoglobulins especially adapted to be transferred from the body fluids to the surface of the tissues with mucous-membrane linings, such as the respiratory and gastrointestinal tracts, and into secretions, such as milk, saliva, tears, and bile. Two four-chain units become joined together and associated with a special protein that enables the IgA molecule to be passed outward through the cells that line the air passages, the gut, and the milk, salivary, tear, and bile ducts.

IgD molecules are present on the surface of most, but not all, B cells early in their development, but little IgD is ever released into the circulation. It is not clear what function IgD performs, though it may play a role in determining whether antigens activate the B cells.

IgE is made by a small proportion of B cells and is present in the blood in low concentrations. Each molecule of IgE consists of one four-chain unit and so has two antigen-binding sites; each of its H chains has an extra constant domain (C_H4), which gives IgE the special property of binding to the surface of basophils, eosinophils, and mast cells (tissue cells that resemble basophils). IgE antibodies help to protect against parasitic infections, but they are also responsible for some allergies (see below *Allergies*).

Most people have fairly constant amounts of immunoglobulin in their blood, which represent the balance between continuous breakdown of these proteins and their new manufacture. There is about four times as much IgG (including its subclasses) as IgA, 10 to 15 times as much as IgM, 300 times as much as IgD, and 30,000 times as much as IgE. This implies that most active B cells make IgG and rather few make IgE.

Normal production of antibody

Part of the normal immunoglobulin production undoubtedly represents the response to antigenic stimulation that happens all the time, but even animals reared in surroundings completely free from microbes and their products make substantial, though lesser, amounts of immunoglobulins. Much of the immunoglobulin therefore must represent the product of all the B cells that are, so to speak, "ticking over" even if not specifically stimulated. It is therefore not surprising that extremely sensitive methods can detect traces of antibodies that react with antigens or antigenic determinants to which an animal has never been exposed but for which cells with receptors are present.

All B cells are potentially able to use any one of the constant-region classes to make up the immunoglobulin they secrete. As noted above, when first stimulated most secrete IgM. Some continue to do so, but others later switch to IgG or IgA or IgE. Memory B cells, which (as

From N.A. Staines, J. Brostoff, and K. James, *Introducing Immunology*; Gower Medical Publishing, 1985

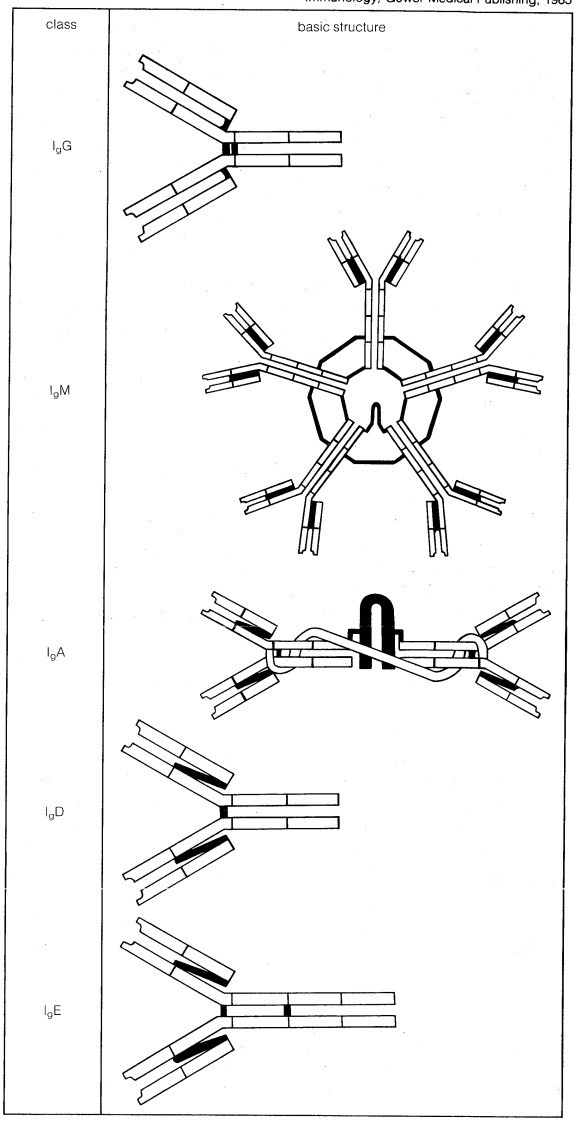


Figure 3: The arrangement of chains in the main classes of immunoglobulins. Links between the chains are represented by dark bars. The additional protein that joins the basic units in IgM and IgA is shown as a dark, inverted U shape; and the protein required for the secretion of IgA is depicted as a twisted ribbon.

explained more fully below) are specialized for responding to repeat infections by a given antigen, make IgG or IgA immediately. What determines the balance among the classes of antibodies is not fully understood. It is influenced by the nature and site of deposition of the antigen (for example, parasites tend to elicit IgE), perhaps through factors released locally by T cells.

T-CELL ANTIGEN RECEPTORS

Since T lymphocytes do not normally shed their antigen receptors, or do so only in very small amounts, understanding of their structure has not come from analysis of isolated receptors but from studies of the DNA that encodes them. These studies show that T-cell antigen receptors consist of two polypeptide chains, each of which, like the chains that make up immunoglobulins, is encoded by a fairly large number of alternative genes that control different parts of the receptor molecule. Also, as with immunoglobulins, these alternative genes are rearranged in any one of a very large number of possible ways to establish the final code for the receptors of any one T cell. The genes are not the same as those that control the manufacture of immunoglobulins, although the resultant molecules have a generally similar structure. So, rather unexpectedly, the body possesses two ways of generating the diversity of antigen receptors that characterize the lymphocytes. As will be mentioned later, this duplication probably resulted from the evolution of both processes from a more primitive and simpler recognition system.

Despite the structural similarities, the receptors on T cells function differently from those on B cells. This functional difference underlies the different roles that B cells and T cells play in the immune system. B cells, as has been explained, secrete antibodies that bind to antigens in the blood and other body fluids. But once an antigen—a virus, for example—succeeds in infecting a body cell, it is safely beyond the reach of the circulating antibodies. T cells, on the other hand, can bind to body cells that harbour pathogens and initiate an immune response against the infected cell. Consequently, T cells must have receptors that can detect infected cells. The way in which they do this is by recognizing foreign molecules that are expressed on the surface of an infected cell in association with normal cell-surface molecules called major histocompatibility complex (MHC) antigens. (In humans the MHC antigens were first discovered on certain leukocytes and are, therefore, often referred to as HLA [human leukocyte group A] antigens. For information on the genetic basis of the HLA antigens, see GENETICS AND HEREDITY, THE PRINCIPLES OF.) The MHC antigens are of two major types: class I antigens, which are present on the surface of virtually all nucleated body cells, and class II antigens, which are found on the surface of most B cells and on some T cells, macrophages, and macrophage-like cells.

Helper T cells and cytotoxic or suppressor T cells

As will be described more fully below, there are two main types of mature T cells. One type consists of helper T cells; the other is made up of cytotoxic or suppressor T cells. Helper T cells recognize foreign antigens associated with MHC class II antigens, and cytotoxic or suppressor T cells generally recognize foreign antigens associated with MHC class I antigens. This recognition is achieved by means of separate molecules on the T cells, closely associated with the specific antigen receptors but able to bind to the self antigens of MHC class I or II. To add even further to the complexity, T-cell antigen receptors are also associated with a third kind of molecule, which is necessary for activating the T cell once its antigen receptor has bound with its specific antigen. The actual arrangement of this receptor complex is not known.

LIFE CYCLE OF T AND B LYMPHOCYTES

When T-cell precursors leave the bone marrow on their way to mature in the thymus, they do not yet express receptors for antigens and are indifferent to stimulation by them. Within the thymus the T cells multiply many times as they pass through a meshwork of thymus cells. In the course of multiplication they acquire antigen receptors and differentiate into the two sorts—helper T cells and cytotoxic or suppressor T cells—mentioned in the previous section. Both kinds look alike, but they can be distinguished by their function and by the presence of special surface proteins. Most of the T cells that have multiplied in the thymus also die there. This seems paradoxical until it is remembered that the random generation of different antigen receptors is bound to produce a large proportion that recognize self antigens—*i.e.*, molecules already present on the body's own constituents—and that mature lymphocytes with such receptors would be liable to attack the body's own tissues. Nearly all such self-reactive T cells die before they leave the thymus, so that those T cells that do emerge are the ones capable of recognizing foreign antigens. These travel via the blood to the lymphoid tissues, where, if suitably stimulated, they can again multiply and take part in immune reactions. The generation of T cells in the thymus proceeds very actively in young animals, but it gradually slows down during adulthood and is much diminished in old age, by which time the thymus has become small and partly atrophied. Cell-mediated immunity persists throughout life, however, because some of the T cells that have emerged from the thymus can survive and function for a very long time.

Elimination of self-reactive T cells

B-cell precursors are continuously generated in the bone marrow throughout life, but as with T-cell generation, the rate diminishes with age. Unless they are stimulated to mature (as described below), the majority of B cells also die, though those that have been stimulated can survive for a long time in the lymphoid tissues. Consequently, there is a continuous supply of new B cells throughout life. Those with antigen receptors capable of recognizing self antigens tend to be eliminated, though less effectively than are self-reactive T cells. As a result, some self-reactive cells are always present in the B-cell population, along with the majority that recognize foreign antigens. The reason why these self-reactive B cells normally do no harm is explained in the following section.

ACTIVATION OF LYMPHOCYTES

The proportion of the total lymphocyte population that can recognize and respond to any given antigen is bound to be small, because most B and T cells bear receptors with other specificities. After a lymphocyte has been activated by a particular antigen, it multiplies to form a clone—that is, a population of identical progeny cells. Each member of the clone bears the same antigen specificity as the original activated lymphocyte. This process, whereby an activated lymphocyte gives rise to a large population of cells targeted on the offending antigen, is known as clonal selection. Because the original population of lymphocytes contains so few members that recognize a particular antigen, immunologists have had to devise special techniques to unravel what goes on when the cells are activated. One such technique is the use of what are called polyclonal mitogens, most of which are plant or bacterial products whose properties were originally discovered by accident. These are molecules that react with some surface component common to all T cells or all B cells, imitating, more or less accurately, activation by an antigen. Polyclonal mitogens can thus activate large numbers of lymphocytes at one time. Another way to study lymphocyte activation is to stimulate B cells or T cells in culture and to select and propagate those that respond to a chosen antigen, thereby obtaining a sufficient population to analyze.

Activation of T cells. Antigen activation of helper T cells only takes place when (1) the foreign antigens are presented at the surface of other cells in association with MHC class II antigens and (2) the T cells have been switched on by interleukin 1 (IL-1), which is secreted by macrophages that have ingested foreign particles or have been activated in other ways (see above *Acute-phase reactions*). As mentioned earlier, MHC II molecules are made by only a few cell types, of which macrophages and B cells are the most relevant. So activation of helper T cells is largely confined to antigens that have first been taken up by macrophages. Since macrophages that have ingested foreign particles and broken down their constituents to smaller fragments return part of these to their cell surface, where MHC molecules are already present, this is an effective way of presenting the antigenic determinants

so as to be recognized by receptors on the helper T cells. The helper T cells then proceed to make another protein, called interleukin 2 (IL-2), as well as receptors that bind it. IL-2, in turn, stimulates the helper T cells to divide and to secrete several more proteins, such as gamma interferons and factors that can trigger B cells to differentiate and secrete immunoglobulins. The overall result is that the number of helper T cells that recognize the specific foreign antigen has been increased, and several T-cell factors, collectively known as lymphokines, have been produced. The lymphokines have other consequences, one of which is that IL-2 allows those cytotoxic or suppressor T cells that also recognize the same antigen to become activated and to multiply. Cytotoxic T cells, in turn, can attack and kill other cells that express the foreign antigen in association with MHC I molecules, which—as explained above—are present on almost all cells. So, for example, cytotoxic T cells can attack target cells that express antigens made by viruses or bacteria growing within them (see below *Cell-mediated immune mechanisms*). Suppressor T cells may be the same as cytotoxic T cells, but they are detected by their ability to suppress the action of B cells or even of helper T cells (perhaps by killing them). Suppressor T cells thus act to damp down the immune response and can sometimes predominate so as to suppress it completely.

Interleukin 2 and T-cell activation

Activation of B cells. A B cell is activated when an antigen binds to its specific receptors. If the B cell is newly formed and immature, the binding of an antigen prevents it from developing any further. As only B cells that recognize foreign antigens survive this stage, this is one way that the immune system prevents reactions against self proteins present in the blood and tissue fluids.

The surface of some antigens—notably the outer coats of many bacteria—consists of polymers (extremely large molecules composed of multiples of simpler chemical units). Such antigens present a large number of identical antigenic determinants and can stimulate the B cells to release antibodies without further ado. Most B cells, however, require the additional stimulation of lymphokines before interaction with the antigen causes them to multiply into a clone of immunoglobulin-secreting cells. These fully activated, antibody-producing cells are called plasma cells, and each can secrete several thousand molecules of immunoglobulin every second and continue to do so for several days. When a specific B cell has been switched on to multiply and make antibody, a large amount of that particular antibody will be released into the circulation. This initial burst of antibody production gradually decreases as the stimulus is removed (*e.g.*, by recovery from infection), but some antibody continues to be present for several months afterward.

Memory cells and long-term immunity

The process just described takes place among the circulating B lymphocytes. But some of the B cells encounter the antigen in the germinal centres—compartments in the lymphoid tissues where few T cells are present—and are activated in a different way. They multiply extensively, especially those with the most effective receptors, but do not go on to secrete antibody at that time. Instead, they remain in the tissues and the circulation for many months or even years. If, with the help of T cells, they encounter the activating antigen again, these B cells rapidly begin to manufacture and release their specific antibody. They behave as though they "remember" the initial contact with the antigen and are therefore termed memory cells. Upon reinfection by a particular microbe, the memory cells trigger a rapid rise in the level of protective antibodies and thus prevent the illness from taking hold.

REGULATION OF IMMUNE RESPONSES

As detailed in the previous sections of this article, specific immunity involves a complex series of interacting processes whereby activated lymphocytes multiply and become effector cells that (1) secrete lymphokines that affect other cells, (2) kill targets bearing the activating antigen, or (3) secrete antibodies that react with the antigen. The biological effects of all of these processes are discussed in later sections.

The way in which the immune system generates the diversity of antigen receptors ensures the production of lymphocytes that can recognize antigenic determinants on almost any molecule—not only those on microbes, but also those on self components and on irrelevant foreign materials that would normally never enter the body. Mechanisms whereby cells bearing receptors for self components are eliminated or suppressed have been discussed, as has the need for antigens to be presented in a special way before they actually trigger immune responses. What has not been mentioned is the fact that specific antigen receptors, being different (if only slightly) between one lymphocyte and another, might themselves be recognized as foreign and stimulate the appropriate T or B cells to respond to them. This does in fact happen. Anti-antibodies develop in the course of an immune response and then anti-anti-antibodies—and probably the process continues. In addition, certain T cells can recognize receptors on B cells and on other T cells and can act either to help or to suppress the cells recognized, thereby regulating how far and for how long the response continues. This regulatory network clearly works rather well in practice, since specific immune responses are sufficient to cope with microbial invasion of the body but fade away, apart from memory cells, once the stimulus has been removed. It is, however, as complicated as the constraints that regulate human societies and about as difficult to analyze in detail.

Role of the central nervous system

There is considerable interest in whether immune responses can be regulated through the central nervous system. Lymphocytes have receptors that can bind certain hormones whose production is controlled by the central nervous system. These hormones can affect immune responses—positively and negatively—both in culture and within the body. There is also evidence that an interleukin produced during immune responses can increase the nervous system's release of messengers that control the secretion of hormones by the adrenal gland. Thus some interplay between the immune system and the nervous system is certainly possible, though the importance of this interaction in regulating immune responses is unclear. Clinical and everyday experience indicates that patients with a "will to live" survive better than those who despair, but whether this higher survival rate is mediated by the impact of the nervous system on the immune system—or indeed, whether it depends upon the functioning of the immune system at all—is unknown.

ANTIBODY-MEDIATED IMMUNE MECHANISMS

Protective attachment to antigens. Many pathogenic microorganisms and toxins can be rendered harmless by the simple attachment of antibodies. For example, some harmful bacteria, such as those that cause diphtheria, tetanus, and plague, release toxins that poison essential body cells. Antibodies, especially IgG, that combine with such toxins neutralize them. Also susceptible to simple antibody attachment are the many infectious microbes—including all viruses, many bacteria, and some protozoans—that live within the body cells; these pathogens bear special molecules by which they must attach themselves to the host cells before they can penetrate and invade. Antibodies that bind to these molecules prevent invasion. Antibody attachment can also immobilize those bacteria and protozoans that swim by means of whiplike flagella or cilia. In all these instances antibodies protect simply by combining with the antigens, although they do not kill or dispose of the microbes. The actual destruction of microbes involves phagocytosis by granulocytes and macrophages, and this requires the participation of the complement system.

Activation of the complement system. Complement comprises 11 proteins that can interact sequentially to produce several biologically important results. The activation of complement makes antigen–antibody complexes attractive to phagocytes; it releases agents that cause inflammation; and, in some cases, it disrupts the outer membranes of antibody-coated cells. These 11 proteins, together with four others whose function is to damp down some of the complement interactions, form the complement system. It is beyond the scope of this article to describe the system in detail, but a brief account is needed because many of the biological effects of antibodies depend on complement activation.

The first point to make is that the complement proteins are made not by lymphocytes but by liver cells and, in many tissues, by macrophages. The second is that the action of complement is nonspecific—*i.e.,* it does not depend upon the specificity of the antigen-binding site of an antibody—and that complement proteins probably evolved before antibodies. Complement functions are similar within many species, and corresponding components from one species can replace those of another. The complement system provides a most ingenious way in which antibodies, whatever their specificity, produce the same biological effects when they combine with antigens.

ctivation
complement by
ntibody

Complement is activated most effectively by IgM and IgG antibodies, which are, as discussed above, the most abundant classes of immunoglobulins. The activation of the complement system by antibodies depends on the H-chain constant regions of the immunoglobulin molecules. When the antigen-binding sites of an immunoglobulin interact with the appropriate antigen, the shape of part of the H-chain constant region is distorted in such a way that it binds to one of the complement components present in the surrounding fluid. This, in turn, rearranges the complement molecule, exposing a catalytic site that can act to split the next complement component in the sequence into two pieces: a small polypeptide fragment and a larger protein part with a catalytic site exposed. This catalyst then acts similarly on the next complement component, which, in turn, acts on a protein called C3—the most abundant and biologically the most important complement component. C3 is split into a small polypeptide and a larger fragment, called C3b, which is the active form of C3.

C3b can do several different things. It can bind to antigen–antibody complexes and confer on them the property of interacting with receptors for C3b present on phagocytic cells. It can also activate the next component in the complement system. This leads to release of yet another small polypeptide and the insertion of a ringlike structure, composed of yet other complement components, into the cell membrane of the invading microbe. This structure causes the cell contents to leak out and the cell to die. Finally, C3b can combine with another protein to form a complex that converts more C3 to C3b. The small polypeptides released during the activation of complement are potent pharmacological agents that (1) cause mast cells and basophils to release histamine, (2) increase the permeability of blood vessels, and (3) attract granulocytes and monocytes.

Thus when a microbe penetrates the body, if antibodies reactive with its surface are already present (or if it activates complement without their help), the complete complement sequence may be activated and the microbe killed by damage to its outer membrane. This mechanism is effective only with bacteria that lack protective coats and with certain large viruses, but it is nevertheless important. Persons who lack C3 and thus cannot complete the later steps in the complement sequence are liable to repeated bacterial infections.

ctivation
f phagocytic cells
y complement

If the microbes are not immediately killed, they nevertheless soon acquire antibody and C3b on their surface, and the small polypeptide fragments released locally divert phagocytic neutrophils and monocytes to the sites of complement activation. The C3b on the microbes binds to the C3b receptors on the phagocytic cells, and the microbes are ingested and usually killed by the digestive enzymes present in the phagocytes. Because granulocytes that have ingested foreign particles release part of their digestive enzymes and activation products into the surrounding fluid, a certain amount of local tissue damage may occur, especially if large numbers of granulocytes have accumulated at the site. This produces local inflammation, which may amount to no more than redness and swelling but can develop into an abscess. Monocytes and macrophages, which move in after the granulocytes, take part in clearing up the debris. These second-line phagocytes then carry the debris to the lymph nodes or other lymphoid tissues, where they present the antigen to lymphocytes and thus may stimulate a new or a renewed immune response. If microbes reach the bloodstream and become coated with antibody and complement there, they are also ingested in a similar fashion by granulocytes or by macrophages in the liver, the spleen, and the bone marrow.

Clearly such a biologically important chain of reactions could do more harm than good if its effects were to spread beyond the site of antigen invasion. Fortunately, the active intermediates at each stage in the complement sequence become rapidly inactivated or destroyed by inhibitors if they fail to initiate the next step. With rare exceptions, this confines the activation to the place where it is needed.

Activation of killer cells. Some cells that bear antigen–antibody complexes do not attract complement; this occurs if the antibody molecules are far apart on the cell surface or are of a class that does not readily activate the complement system. Other cells have outer membranes that are so tough or can be repaired so quickly that the cells are impermeable to activated complement. Still others are so large that phagocytes cannot ingest them. Such cells can be attacked by so-called killer (K) cells present in the blood and lymphoid tissues. K cells look like lymphocytes but are not typical T or B cells; they may, in fact, be the same as the NK cells mentioned earlier as contributing to nonspecific, innate immunity. K cells have receptors that bind to the H chains of IgG at the ends away from the antigen-binding site. These parts of the H chains are exposed on a target cell with antibody attached to it, and so K cells can bind quite tightly to them. Once bound, K cells insert something—its nature is uncertain—that causes the cell to swell and burst. K cells do not harm bacteria, but they play a role in destroying cells infected by viruses.

Role of K cells

Other antibody-mediated mechanisms. The protection afforded by IgA antibodies, which, it will be recalled, are transported to the surface of mucous-membrane-lined passages, is somewhat different. Complement activation is not involved since there are no complement proteins on the lining of the gut or the respiratory tract. Here protection consists largely of IgA combining with microbes to prevent them from entering the cells of the lining. The bound microbes are then swept out of the body.

IgE antibodies also invoke unique mechanisms. As stated earlier, most IgE molecules are bound to special receptors on mast cells and on basophil and eosinophil granulocytes. When antigens bind to IgE antibodies on these cells, the consequence is not to cause ingestion of the antigens but to trigger the release of the cells' granules and their contents. In the case of mast cells and basophils, this results in the sudden increase in permeability of the local blood vessels; the adhesion and activation of platelets (blood cell fragments that trigger clotting), which release their own active agents; the contraction of smooth muscle in the gut or in the respiratory tubes; and the secretion of fluids—all of which tend to dislodge large multicellular parasites such as hookworms. In the case of eosinophils, which plaster themselves onto such parasites as the flatworms that cause schistosomiasis, the granule contents are particularly effective at destroying the parasite's tough, protective skin. Therefore, IgE antibodies—although they can be a nuisance when they react with otherwise harmless antigens, as discussed below in the section on *Allergies*—appear to have a special protective role against the larger parasites.

Transfer of antibodies from mother to offspring. A newborn mammal has had no opportunity to develop protective antibodies on its own, unless, as happens very rarely, it became infected within the uterus. Yet it is born into an environment similar to its mother's, with all the potential microbial invaders to which she is exposed. Although it possesses the mechanisms of innate immunity, it has none of its mother's lymphocytes. The placenta strictly excludes the maternal lymphocytes from crossing into the fetus, where they would treat the fetal tissues as foreign antigens and thereby cause a reaction similar to the rejection of an incompatible organ transplant (see TRANSPLANTS, ORGAN AND TISSUE).

What is transferred across the placenta in many species is a fair sample of the mother's antibodies. How this happens depends upon the structure of the placenta, which varies with the species. In humans the mother's IgG antibodies (but not those of the other immunoglobulin classes) are actively transported across the placenta throughout

the second two-thirds of pregnancy. In many rodents a similar transfer occurs, but mainly by crossing through the yolk sac.

In horses and cattle, which have more layers of cells in their placenta, no antibodies are transferred during fetal life, and the newborn is unprotected. There is, however, a second mechanism that makes up for this deficiency. The early milk (colostrum) is very rich in antibodies, mainly IgA but also some IgM and IgG, and during the first few days of life the newborn mammal can absorb these proteins intact from the digestive tract directly into the bloodstream. Drinking colostrum is therefore essential for newborn horses and cattle and to a somewhat lesser extent for other mammals. The capacity of the digestive tract to absorb intact proteins must not last beyond one or two weeks, since once foods other than milk are ingested the proteins and other antigens in them would also be absorbed intact and could act as immunogens to which the growing animal would become allergic (see below *Allergies*). IgA in milk is, however, rather resistant to digestion and can function within the gut even though intact absorption into the bloodstream has ended. Human colostrum is also rich in maternal antibodies, but they are less essential and are absorbed intact only during the first day or so after birth.

After a newborn has received its supply of maternal antibodies, the child is as fully protected as its mother. This means, of course, that if the mother has not developed immunity to a particular pathogen, the newborn will likewise be unprotected. For this reason, a physician may recommend that a prospective mother receive immunizations against tetanus and certain other disorders. (The active immunization of pregnant women against certain viral diseases, such as rubella [German measles], must be avoided, however, because the immunizing agent can cross the placenta and produce severe fetal complications.)

As important as the passively transferred maternal antibodies are, their effects are only temporary. The maternal antibodies in the blood become diluted as the animal grows; moreover, they gradually succumb to normal metabolic breakdown. Because the active development of acquired immunity is a slow and gradual process, young mammals actually become more susceptible to infection during their early stages of growth than they are immediately after birth. This is illustrated in Figure 4.

Erythroblastosis fetalis and Rh antibody

Occasionally the transfer of maternal antibodies during fetal life can have harmful consequences. A well-known example of this is erythroblastosis fetalis, or hemolytic disease of the newborn, a disorder in which maternal antibodies destroy the child's red blood cells during late pregnancy and shortly after birth. The most severe form of erythroblastosis fetalis is Rh hemolytic disease, which develops when: (1) The fetus is Rh-positive; that is, its red blood cells carry an antigen known as the Rh factor. (2) The mother is Rh-negative, which is to say her red blood cells lack the Rh factor. (3) The mother's immune system has been previously activated against the Rh antigen; this usually is the result of exposure to fetal cells during the birth of an earlier Rh-positive baby. Rh hemolytic disease can be prevented by giving the mother injections of anti-Rh antibody shortly after the birth of an Rh-positive child. This antibody destroys any Rh-positive fetal cells in the maternal circulation, thereby preventing the activation of the mother's immune system.

From O. Orlandini, A. Sass-Kortsak, and J.H. Ebbs, "Serum Gamma Globulin Levels in Normal Infants," *Pediatrics*, copyright 1955, vol. 16, p. 575

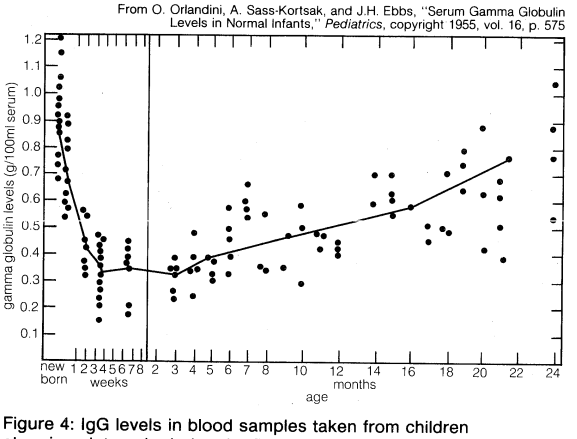


Figure 4: IgG levels in blood samples taken from children at various intervals during the first two years of life. The declining levels during the first weeks reflect the disappearance of immunoglobulins transferred from the mother before birth. The gradually increasing levels from the age of three months result from the spontaneous production of the child's own antibodies. Each point represents one measurement; the line is drawn between the means.

CELL-MEDIATED IMMUNE MECHANISMS

In addition to their importance in cooperating with B cells that secrete specific antibodies, T cells have important, separate roles in protecting against antigens that have escaped or bypassed antibody defenses. Immunologists have long recognized that antibodies do not necessarily protect against viral infections, because many viruses can spread directly from cell to cell and thus avoid encountering antibodies in the bloodstream. It is also known that persons who fail to make antibodies are very susceptible to bacterial infections but are not unduly liable to viral infections. Protection in these cases results from cell-mediated immunity. It depends upon the ability to destroy and dispose of body cells in which viruses or other intracellular parasites (such as the bacteria that cause tuberculosis and leprosy) are actively growing, as well as to dispose of the microbes themselves.

T-cell protection against viral infection

Cell-mediated immunity has two components. One depends upon the elaboration of lymphokines by helper T cells that have interacted with the appropriate antigen. In particular the gamma interferon produced by the helper T cells greatly increases the ability of macrophages to kill ingested microbes; this can tip the balance against microbes that otherwise resist killing. Gamma interferon also stimulates the NK cells mentioned in the section on nonspecific immunity. The second component of cell-mediated immunity depends upon cytotoxic T cells. These attach themselves by their receptors to cells whose surface expresses appropriate antigens (notably ones made by developing viruses) and somehow damage the infected cells enough to kill them. After a sufficient number of antigen-specific helper and cytotoxic T cells are available and the cells that act nonspecifically have been stimulated, most viral infections are rapidly overcome.

IMMUNITY AGAINST CANCER

The role of the immune system in protecting against cancer has not been fully explained. One widely held theory, called the immune surveillance hypothesis, postulates that cancer cells arise frequently but are rapidly destroyed by immune responses. This idea can be traced back to the early years of the 20th century, when Ehrlich pointed out that the enormous multiplication and differentiation of cells during prenatal life must afford many opportunities for aberrant cells to appear and grow. He speculated that immune mechanisms must eliminate such cells, since there is normally no sign of them at birth. That such mechanisms continue to function throughout life, weeding out newly arisen cancer cells, forms the basis of the immune surveillance hypothesis.

Destruction of cancer cells

Experiments have shown that cancer cells can be killed by immune processes. For example, cancer cells taken from one inbred strain of mouse or rat will grow readily when transplanted into another animal of the same inbred strain, but they are usually unable to grow for more than a few days when transplanted into mice or rats of a different inbred strain. The mechanism of rejection is the same as that responsible for the rejection of tissue transplants between different strains; it depends upon a cell-mediated immune response against the MHC antigens present on the cancer cells (see TRANSPLANTS, ORGAN AND TISSUE).

While the experiment described above shows that immune responses can kill cancer cells, it does not explain how this might happen—as the immune surveillance hypothesis says it should—when the cancer cells arise within an individual and thus bear that individual's MHC antigens. Other experiments demonstrate that many cancer cells carry unique non-MHC antigens while lacking one

or more of the normal antigens, including MHC-I antigens. These cancers might therefore be expected to evoke an immune response, and frequently one can be detected. Antibodies specific for the cancer cells are often present in the blood, usually in low concentrations, and lymph nodes "downstream" from the cancer site usually contain some lymphocytes that recognize the cancer-cell antigens.

Despite the demonstrable immune responses, cancers do develop. Experiments with animal tumours have provided several possible explanations for this, any or all of which may apply to a given cancer. The simplest explanation is that the cancer grows so fast or in such a thick mass that the cytotoxic T cells or NK cells cannot get at most of the cancer cells. Another explanation is that cancer cells that have lost their MHC-I antigens can no longer be killed by cytotoxic T cells. Experiments also show that some cancer cells secrete factors that inhibit NK cells as well as interfere with the immune response; that cancer-induced immune responses often entail a predominance of suppressor T cells over helper T cells; that some cancer cells can shed their antigens, which then block the receptors on the cytotoxic T cells and thereby prevent the lymphocytes from attacking the cancer cells; and that cancer cells often resemble undifferentiated cells present during embryonic life, to which the body would not be expected to become immune.

The fact that cancers nevertheless develop does not prove that immune responses against them are useless. Transplant patients and others treated with immunosuppressive drugs over long periods of time are more liable to develop certain kinds of cancer, especially those caused by viruses. The incidence of cancer also increases greatly in old age, when immune responses are known to decline; however, the chance of having accumulated sufficient genetic mutations to cause cancer also increases sharply with age.

Although injections of interferon and other efforts to stimulate immunity in cancer patients have not been notably successful, much research has been devoted to developing effective immunotherapies against cancer. Another avenue of research has focused on finding ways to immunize healthy persons against cancer. Protection against implanted tumour cells has been induced in experimental animals by immunization with killed cancer cells or antigens from them. But this requires the use of risky procedures to increase the immune response, and it is not therefore acceptable as preventive therapy in humans.

PROPHYLACTIC IMMUNIZATION

Prophylactic immunization refers to the artificial establishment of specific immunity, a technique that has significantly reduced suffering and death from a variety of infectious diseases. There are two types of prophylactic immunization: passive immunization, in which protection is conferred by preformed antibodies; and active immunization, in which protection results from the administration of vaccines that stimulate a specific immune response.

Passive immunization. The administration of preformed antibodies can provide lifesaving assistance in combatting poisons or infections that produce dangerous effects so rapidly that the victim does not have time to develop an immune response spontaneously. This situation may arise with victims of poisonous snakebites or botulism, as well as those in whom such infections as diphtheria, tetanus, or gas gangrene have progressed to the point where bacterial toxins have been absorbed. It is also the case with bites from a rabid animal, although active immunization is also begun at the same time since the spread of the rabies infection to the central nervous system is relatively slow. Physicians also use passive immunizations to protect temporarily persons traveling to countries where hepatitis B is prevalent and to provide antibodies to persons who suffer from B-cell deficiencies and are therefore unable to make antibodies for themselves (see below *Immune deficiencies*). As discussed earlier, physicians use passive immunizations of anti-Rh antibody to prevent erythroblastosis fetalis.

Manufacture of antiserum

Protective immunoglobulins—primarily of the IgG class—can be prepared from the blood of humans or other species (*e.g.*, horses or rabbits) that have already developed specific immunity against the relevant antigens. These preparations are known as antiserums. Human IgG is only slowly broken down in the recipient's body, the concentration falling by about one-half every three weeks, so that effective amounts of antibody can be present for two or three months. IgG from other species is far more likely to provoke an immune response that will eliminate the antibody and can lead to serum sickness (see below *Allergies*), so human antiserum is chosen whenever it is available.

Active immunization. Active immunization aims to ensure that a sufficient supply of antibodies or T and B cells that react against a potential infectious agent are present in the body before infection occurs. So primed, the immune system either can prevent the pathogen from establishing itself or can rapidly mobilize the various protective mechanisms described above to abort the infection in its earliest stages.

Manufacture of vaccine

The vaccines used to provide active immunization need not contain living microbes. What matters is that they include the antigens important in evoking a protective response and that those antigens are administered in a harmless form and are sufficient in amount and persistence to produce an immune response similar to the natural infection. Bacterial toxins, such as tetanus or diphtheria, can be rendered harmless by treatment with formaldehyde without affecting their ability to act as immunogens. They are usually administered adsorbed onto an inorganic gel, which helps to retain them in macrophages, and they elicit effective, long-lasting immunity against the toxins themselves. In other instances, when immunization against several antigenic determinants is desirable or the important antigenic component is not known, it may be better to use the microbes themselves killed in such a way as not to alter them significantly. Such killed vaccines are used to immunize against typhoid, whooping cough, plague, and influenza, for example. In other cases researchers have developed attenuated (*i.e.*, weakened) strains of bacteria or viruses, which cause an infection but do not produce the full disease, multiply only to a limited extent in the body, and never revert to the virulent form. The use of such live microbes provides the most effective prophylaxis of all, since they truly imitate a mild form of the natural infection. Such are the vaccines for smallpox, yellow fever, poliomyelitis (oral vaccine), measles, rubella, and tuberculosis. Although attenuated so far as normal persons are concerned, live vaccines may cause the full disease in persons who have an immune deficiency.

When the same diseases can be caused by many strains of microbes, each characterized by different antigenic determinants, neither natural infection nor prophylactic immunization with any one strain protects against infection by the others. For example, it is impractical to immunize against the many viruses that cause the common cold or the numerous streptococci that produce tonsillitis. On the other hand, although there are more than 60 different strains of pneumococci that can cause bacterial pneumonia, some of the strains are much more common than others. Consequently a vaccine containing antigens from up to 14 of the most common strains has proved useful in protecting persons at special risk.

Active immunization is the most effective and cheapest way of protecting against an infectious disease. Furthermore, if 95 percent or more of the population at risk is protected and if humans are the only reservoir of infection, active immunization can even lead to the eradication of the infectious agent. This has been achieved worldwide in the case of smallpox.

PRODUCTION OF MONOCLONAL ANTIBODIES

Before turning to the disorders of the immune system, it is appropriate to discuss the artificial production of monoclonal antibodies, a technique that has emerged as one of the most important facets of biotechnology during the last quarter of the 20th century.

It was pointed out above that upon activation by an antigen, a circulating B cell multiplies to form a clone of plasma cells, each secreting the identical immunoglobulin. It is such an immunoglobulin—derived from the descendants of a single B cell—that constitutes a monoclonal

Problem of isolating specialized antibody

antibody. The antibody response to a natural infection or an active immunization, however, is polyclonal. In other words, it involves many B-cell clones, each of which recognizes a different antigenic determinant and secretes a different immunoglobulin. Thus the blood serum of an immunized person or animal normally contains a mixture of antibodies, all capable of combining with the same antigen but with a variety of specificities. The isolation of an appreciable quantity of a particular (*i.e.,* monoclonal) antibody from this polyclonal mixture is extremely difficult.

There is, however, a condition in which the blood serum may contain an astonishingly high concentration of a single immunoglobulin. This results from multiple myeloma, a type of cancer in which a single B cell proliferates to form a tumorous clone of antibody-secreting cells (see below *Cancers of the lymphocytes*). The immunoglobulins made by myelomas are monoclonal, but although they must be antibodies capable of combining with some antigen, there is usually nothing to indicate what this antigen might be. Myelomas are not uncommon in species other than humans, especially in mice and rats, and they can be made to occur frequently in certain laboratory strains by injecting the experimental animal with mineral oil, which acts as a mild irritant and causes B-cell proliferation. Testing of immunoglobulins from many mouse myelomas for their capacity to bind to a very wide variety of antigens has indeed shown that some of them have recognizable antibody activity. Researchers can propagate individual myelomas by growing them in other animals of the same strain or by culturing them outside the body. Furthermore, like other lines of cancer cells, cultured myelomas can multiply indefinitely. (This stands in contrast to normal cell lines, which die out after a certain number of cell divisions.) Thus the propagation of myelomas has enabled immunologists to reap large quantities of monoclonal antibody and has been invaluable in studying the basic nature of immunoglobulins. But if an immunologist wishes to obtain large amounts of a particular antibody—say the anti-Rh antibody—the induction of myelomas is useless, for it has proved impossible to specify beforehand what antibody will be secreted by any given myeloma.

Occasionally, however, a cultured myeloma cell line continues to grow well but loses its ability to secrete immunoglobulin. In 1975 immunologists Georges Köhler and César Milstein discovered that cultured myeloma cells could be made to fuse with normal B cells from the spleen of an immunized mouse. The fusion of a myeloma cell from a line that has lost the ability to secrete immunoglobulin with a B cell known to secrete a particular antibody results in a remarkable one that makes the antibody characteristic of its B-cell component but retains the capacity of its myeloma component to multiply indefinitely. Thus it is possible to obtain hybrid cells, termed hybridomas, that grow like a myeloma but make a chosen, identifiable monoclonal antibody. Since only some of the myeloma cells fuse with the spleen B cells, and only a few of these may be making the desired antibody, researchers have devised means to select the wanted hybridomas from the rest of the cells in the mixture and to propagate them. (To learn more about these techniques, see the *Bibliography.*)

Manufacture of specialized hybrid cells

Thanks to hybridomas, researchers can obtain monoclonal antibodies that recognize individual antigenic sites on almost any molecule, ranging from drugs and hormones to microbial antigens and cell receptors. The exquisite specificity of monoclonal antibodies and their availability in quantity has made it possible to devise sensitive assays for an enormous range of biologically important substances and to distinguish cells from one another by identifying previously unknown marker molecules on their surfaces. For example, monoclonal antibodies that react with cancer antigens can be used to identify cancer cells in tissue samples. Moreover, if short-lived radioactive atoms are added to these antibodies and they are then administered in tiny quantities to a patient, they become attached exclusively to the cancer tissue. By means of instruments that detect the radioactivity, physicians can locate the cancerous sites without surgical intervention. Monoclonal antibodies have also been used experimentally to deliver cytotoxic drugs or radiation to cancer cells.

Although the preparation of monoclonal antibodies with hybridomas derived from rat or mouse cells has become routine practice, it has not proved so easy to obtain human hybridomas. This is partly because most human myeloma cells do not grow well in culture, and those that do so have not produced stable hybridomas. If, however, human B cells isolated from blood are infected by the Epstein–Barr virus (the agent that causes infectious mononucleosis), they can be propagated in culture and continue to secrete immunoglobulin. Very few of them are likely to be making an antibody with a desired specificity, even in a subject who has been immunized; but in some instances immunologists have succeeded in identifying and selecting those cells that secrete the wanted immunoglobulin. These can be grown in culture as single clones that secrete a monoclonal antibody. Researchers have used this process to obtain human monoclonal antibodies against the Rh antigen.

Disorders involving the immune system

IMMUNE DEFICIENCIES

There are several ways in which the protective mechanisms outlined above may fail. Some are inborn, due to genetic defects in the development of one or more of the cells involved in immune responses. Others result from infectious agents that damage essential immune cells. Still others are due to poisons or to drugs administered accidentally or with the intention of curing or ameliorating other diseases. In yet other cases, the immune deficiency stems from inadequate nutrition.

Genetic deficiencies

Some infants never develop a thymus, and consequently they cannot produce mature T lymphocytes. In others the B lymphocytes fail to secrete immunoglobulins; some infants are born with combined immune deficiency, in which the functioning of both types of lymphocytes is impaired. Other inborn deficiencies occur because the macrophages or granulocytes lack one or more of the enzymes needed to destroy ingested microbes. All these conditions are rare, and many victims can be maintained in reasonable health by drastic procedures, such as a thymus or bone-marrow transplant to restore the defective cells, or by regular administration of pooled immunoglobulins prepared from the blood of normal people.

Damage to lymphocytes by viruses is more common but usually transient. During infectious mononucleosis, for example, the Epstein–Barr virus infects B cells, causing them to express viral antigens. T cells that react against these antigens then attack the B cells, and a temporary deficiency in the production of new antibodies lasts until the overt viral infection has been overcome. Since antibodies already present in the blood are only slowly broken down, failure to make new ones is only important if the infection persists for a long time, as it sometimes does. A more serious viral infection is that due to the agent responsible for AIDS (acquired immune deficiency syndrome), which selectively infects helper T cells and prevents them from producing lymphokines and from functioning in cell-mediated immunity. Persons with AIDS may be unable to overcome infections by a variety of microbes that are easily disposed of by normal persons. Severe infections by certain parasites, such as trypanosomes, also cause immune deficiency, as do some forms of cancer, but it is uncertain how this comes about.

Deficiencies caused by drug therapy

Probably the most common cause of immune deficiency in countries with advanced medical services is the use of powerful drugs to treat cancers. These drugs work by inhibiting the multiplication of rapidly dividing cells. Although the drugs are chosen so far as possible to act selectively on the cancer cells, they are also liable to interfere—even if to a lesser extent—with the generation and multiplication of the cells involved in immune responses. Prolonged or intensive treatment commonly impairs immune responses to some degree. Although this immune impairment is reversible, the physician must seek a balance between intentional damage to the cancer cells and unintentional damage to the immune system.

Medically induced suppression of the immune system also occurs when powerful drugs, which are designed

specifically to interfere with the development of T and B cells, are used to prevent the rejection of organ or bone-marrow transplants or to damp down serious autoimmune responses (see TRANSPLANTS, ORGAN AND TISSUE). Although use of such drugs has greatly improved the success of transplantation, it also leaves the patient highly susceptible to microbial infections. Fortunately, most such infections can be treated with antibiotics, but the immunosuppressive drugs have to be used with great care and for as short a period as possible.

In countries where the diet, especially that of growing children, is grossly inadequate in respect to protein intake, severe malnutrition ranks as an important cause of immune deficiency. Antibody responses and cell-mediated immunity are seriously impaired, probably due to atrophy of the thymus and the consequent deficiency of helper T cells. This renders the children particularly susceptible to measles and diarrheal diseases. Fortunately, the thymus and the rest of the immune system can recover completely if adequate nutrition is restored.

ALLERGIES

As explained earlier, the immune system must be able to recognize and respond to almost any foreign molecule, since it cannot foretell which molecules will be characteristic of potentially infective agents and which will not. Consequently, an immune response can be induced by materials that have nothing to do with infection, and the mechanisms brought into play, though beneficial for eliminating microbes, are not necessarily beneficial when otherwise innocuous substances are concerned. Furthermore, even initially protective mechanisms can cause secondary disorders when they operate on too great a scale or for longer than necessary, thereby damaging tissues remote from the infection. The terms allergy and hypersensitivity are commonly used to describe inappropriate immune responses that occur when a subject is reexposed to substances that were harmless when encountered for the first time.

The manifestations of a particular allergic reaction depend upon which of the immune mechanisms predominates in the response. Based on this criterion, immunologists recognize four types of hypersensitivity reactions. Types I, II, and III involve antibody-mediated mechanisms and are of rapid onset. Type IV, on the other hand, stems from cell-mediated mechanisms and has a delayed onset. Although this categorization is useful, it should be noted that an allergy can involve a combination of hypersensitivity reactions.

IgE antibody and allergic reaction

Type I hypersensitivity. This type of reaction, also known as atopic or anaphylactic hypersensitivity, occurs when the antigen reacts with IgE antibodies bound to tissue mast cells. Introduction of the antigen links the IgE molecules together and triggers the extrusion of granules from the mast cells and the release of the granular contents into the surrounding fluids. The granules contain histamine, which dilates small blood vessels and causes smooth muscle in the bronchial tubes of the lung to constrict; heparin, which prevents blood coagulation; enzymes that break down proteins; agents that attract eosinophil and neutrophil granulocytes; and a chemical that stimulates platelets to adhere to blood vessel walls and to release serotonin, which constricts arteries. In addition, the stimulated mast cells make chemicals (prostaglandins and leukotrienes) that have potent local effects: they cause capillary blood vessels to leak, smooth muscles to contract, granulocytes to move more actively, and platelets to become sticky.

The overall result of the antigen-IgE-mast cell reaction is an acute inflammation marked by local seepage of fluid from and dilatation of the blood vessels, followed by ingress of granulocytes. As explained earlier, this inflammatory reaction can be a useful local protective mechanism. If, however, it is triggered by an otherwise innocuous antigen entering the eyes and nose, it results in swelling and redness of the linings of the eyelids and nasal passages, secretion of tears and mucus, and sneezing—the typical symptoms of hay fever. If the antigen penetrates the lungs, not only do the linings of the bronchial tubes become swollen and secrete mucus, but the muscle in their walls contracts and the tubes are narrowed, making expiration particularly difficult. These are the symptoms of acute asthma. If the antigen is injected beneath the skin—for example, by the sting of an insect or in the course of some medical procedure—the swelling may be extensive though still local. If, however, the injected antigen gets into the bloodstream and interacts with basophils in the blood as well as with mast cells in the tissues, the release of active agents can cause hives (urticaria, or nettle rash) all over the body. Or, if the reaction is more extensive and sudden, it can produce a potentially fatal condition known as anaphylaxis, or anaphylactic shock, which is characterized by a profound and prolonged fall in blood pressure accompanied by difficulty in breathing. If the antigen enters through the gut, the consequences can include painful intestinal spasms and vomiting. Since, moreover, local reaction with mast cells increases the permeability of the mucosa of the gut, in many cases the antigen enters the bloodstream and also produces hives. This sort of rash is accompanied by severe itching, which is a characteristic effect of histamine on nerve endings in the skin.

Anaphylactic shock

Most people are not unduly susceptible to hay fever or to asthma. Those who are—about 10 percent of the population—are sometimes described as atopic. They have an increased tendency to make IgE antibodies. This tendency runs in families, though there is no single gene responsible as there is in such hereditary diseases as hemophilia. Although many innocuous antigens can stimulate a little IgE antibody in the atopic individual, some antigens are much more likely to do so than others, especially if they are repeatedly absorbed in very small amounts through mucous surfaces. Such antigens are often termed allergens. They are usually polypeptides with attached carbohydrate groups and are resistant to drying, but no special characteristic is known that clearly distinguishes them from all other antigens. Allergens are present in many types of pollen (which accounts for the seasonal incidence of hay fever), in fungal spores, in animal dander and feathers, in plant seeds (especially when finely ground) and berries, and in what is called house dust. The main allergen in house dust has been identified as the excreta of mites that live on skin scales; other mites (those that live in flour, for example) also excrete potent allergens. This list is far from exhaustive. Sensitivities to chocolate, egg whites, oranges, or even cow's milk are not uncommon.

The amount of allergen needed to trigger an acute type I hypersensitivity reaction in a sensitive person is very small: less than one milligram can produce fatal anaphylaxis if it enters the bloodstream. Medical personnel should inquire about any history of hypersensitivity before administering drugs by injection, and if necessary they should inject a test dose into (rather than through) the skin to ensure that hypersensitivity is absent. In any case, a suitable remedy should be at hand.

Another feature of type I hypersensitivity reactions is that once the immediate local reaction to the allergen has taken its course, there may occur an influx of more basophil granulocytes at the site. If the allergen is still present, a more prolonged form of the same reaction—lasting for hours rather than minutes—may supervene. This is a feature of asthmatic attacks in some subjects, in whom repeated episodes also lead to increased sensitivity of the air passages to the constrictive action of histamine. If such persons can escape exposure to the allergen for several weeks, subsequent exposure causes much less severe attacks. A prolonged IgE-induced reaction also causes atopic dermatitis, a skin condition characterized by persistent itching and scaly red patches. These often develop at sites where the skin is bent, such as the elbows and knees. The persistence is due to the influx of mast cells stimulated by the continued presence of the allergen, which is often something absorbed from the gut, such as cow's milk or egg protein.

Several drugs are available that mitigate the effects of IgE-induced allergic reactions. Some, such as sodium chromoglycate and aminophylline, prevent mast-cell granules from being discharged; others, which have effects similar to adrenaline, counteract the effects of histamine. For

Anti-histamines, cortico-steroids, and other drugs

treatment of asthma and severe hay fever, these drugs are best administered by inhalation. The effects of histamine can also be blocked (rather than counteracted) by anti-histamine agents that compete with histamine for binding sites on the target cells. Antihistamines are used to control mild hay fever and such skin manifestations as hives, but they tend to make people sleepy. Corticosteroid drugs can help control persistent asthma or dermatitis, probably by diminishing the inflammatory influx of granulocytes, but long-continued administration can produce dangerous side effects and should be avoided.

Sensitivity to allergens often diminishes with time. One explanation is that increasing amounts of IgG antibodies are produced, which preferentially combine with the allergen and so prevent it from reacting with the cell-bound IgE. This is the rationale for desensitization treatment, in which small amounts of the allergen are injected beneath the skin in gradually increasing quantities over a period of several weeks, so as to stimulate IgG antibodies. The method is often successful in diminishing hypersensitivity to a tolerable level or even abolishing it. However, increased IgG production may not be the complete explanation. The capacity to make IgE antibodies depends upon the cooperation of helper T cells, and they in turn are regulated by suppressor T cells. There is suggestive evidence that atopic individuals are deficient in suppressor T cells whose function is specifically to depress the B cells that produce IgE and that desensitization treatment may overcome this deficiency.

Type II hypersensitivity. Allergic reactions of this type occur when cells within the body are destroyed by antibodies, with or without activation of the complement system. By contrast with type I reactions, in which antigens interact with cell-bound immunoglobulins, type II reactions involve the interaction of circulating immunoglobulins with cell-bound antigens. Type II reactions (also known as cytotoxic reactions) only rarely result from the introduction of innocuous antigens. More commonly, they develop because antibodies have formed against body cells that have been infected by microbes (and thus present microbial antigenic determinants) or because antibodies have been produced that attack the body's cells themselves. This latter process underlies a number of autoimmune diseases, including autoimmune hemolytic anemia, myasthenia gravis, and Goodpasture's syndrome, all of which are discussed later in this article (see *Autoimmune disorders*).

Reactions against blood transfusions

Type II reactions also occur after incompatible blood transfusion, when red blood cells are transfused into a person who already has antibodies against them (either naturally or as a result of previous transfusions). Such transfusions are largely avoidable (see BLOOD: *Uses of blood grouping*), but when they do occur the effects vary according to the class of antibodies involved. If these activate the complete complement system, the red cells are rapidly hemolyzed (made to burst) and the hemoglobin in them is released into the bloodstream. In small amounts it is mopped up by a special protein called hemopexin, but in large amounts it is excreted through the kidneys and can damage the kidney tubules. If activation of complement only goes part of the way (to the C3 stage), the red cells are taken up and destroyed by granulocytes and macrophages, mainly in the liver and spleen. The heme pigment from the hemoglobin is converted to the pigment bilirubin, which accumulates in the blood and makes the subject appear jaundiced. Erythroblastosis fetalis, which was described above, also stems from a type II reaction; and the accumulation of bilirubin in the newborn baby can damage the brain.

Type III hypersensitivity. Type III, or immune-complex, reactions include the various forms of damage caused by the activation of complement in response to antigen–antibody complexes that have become deposited in the tissues or in the walls of the blood vessels. For significant damage to occur, the antigen must be present for longer periods and in greater amounts than those required to cause type I hypersensitivity. The damage results from the various consequences of complement activation already described, especially the influx and activation of neutrophils and macrophages. These scavenger cells release enzymes (which destroy tissues locally) and interleukin 1 (which, among other things, causes fever), secreting these chemicals faster than they can be neutralized or dispersed.

Type III reactions caused by repeated exposure

There are many examples of type III hypersensitivity due to the deposition of antigens in tissues where preformed IgG antibodies are present. For instance, when insects bite they inject their saliva through the skin. The first few times this occurs there may be little reaction; but later bites by the same species of insect give rise to greater and longer-lasting inflammation, owing to the supervening hypersensitivity reaction. More important are those reactions caused by the inhalation of antigens into the lung: for example, farmer's lung, due to fungal spores in moldy hay; pigeon-fancier's lung, due to proteins in powdery pigeon dung; humidifier fever, due to otherwise harmless protozoans that flourish in air-conditioning plants and become dispersed in fine droplets in the air of offices. In each case the farmer, the pigeon fancier, or the office worker will have IgG antibodies against the agent in question in their blood. Inhalation of the antigen results in tightness of the chest, fever, and malaise; these symptoms usually pass in a day or two but recur with reexposure. Permanent damage is rare unless the victims are exposed repeatedly. Some occupational diseases of workers who handle cotton, sugarcane, or coffee waste in warm countries have a similar cause, with the sensitizing antigen usually coming from fungi that grow on the waste rather than the waste itself. The effective treatment is, of course, to prevent further exposure.

The consequences of interaction of antigens with antibody within the bloodstream vary according to whether the complexes formed are large, in which case they are mostly trapped and removed by macrophages in the liver, spleen, and bone marrow, or small, in which case they remain in the circulation. Large complexes occur when more than enough antibody is present to bind to all the antigen molecules, so that these become linked together by the antibody molecules with their multiple binding sites. When the amount of antibody in relation to the antigen is only enough to form small complexes, which can nevertheless activate complement, these tend to settle in the narrow capillary vessels of the synovial tissue (the lining of joint cavities), the kidney, the skin, or, less commonly, the brain or the mesentery of the gut. The activation of complement—which leads to increased permeability of the blood vessels, release of histamine, stickiness of platelets, and attraction of granulocytes and macrophages—becomes more important when the antigen–antibody complexes are deposited in blood vessels than when they are deposited in the tissues outside the capillaries. The symptoms, depending on where the damage occurs, are swollen, painful joints; a raised skin rash; nephritis (kidney damage, causing blood proteins and even red blood cells to leak into the urine); diminished blood flow to the brain; or gut spasms.

Serum sickness

The type of allergy described in the preceding paragraph was first recognized as serum sickness, a condition that often occurred after animal antiserum had been injected into a patient to destroy diphtheria or tetanus toxins. While still circulating in the blood, the foreign proteins in the antiserum induced antibodies, and some or all of the symptoms described above developed in many subjects. Serum sickness is rare nowadays, but similar symptoms can develop in people who develop sensitivity to penicillin or to certain other drugs, such as sulfonamides. In such cases, the drug combines with the subject's blood proteins so that they now have a new antigenic determinant on them, to which antibodies are formed.

The formation of troublesome antigen–antibody complexes in the blood can also result from subacute bacterial endocarditis, a chronic infection of damaged heart valves. The infectious agent is often *Streptococcus viridans,* normally a harmless inhabitant of the mouth. The bacteria in the heart become covered with a layer of fibrin, which protects them from destruction by granulocytes; but they continue to release antigens into the circulation. These can combine with preformed antibodies to form immune complexes that can cause symptoms resembling those of serum sickness. Treatment involves eradication of the heart infection by a prolonged course of antibiotics.

Virtually all autoimmune diseases, in which antibodies are made against self antigens, lead to the appearance of immune complexes in the blood. In most cases, the amount is too small to cause type III hypersensitivity reactions. Two exceptions are rheumatoid arthritis and systemic lupus erythematosus, both of which involve type III reactions. These disorders are discussed below under *Autoimmune disorders.*

Type IV hypersensitivity. As mentioned above, type IV hypersensitivity is cell-mediated; in other words, it does not involve the participation of antibodies but is due primarily to the interaction of T cells with antigens. Reactions of this kind depend upon the presence in the circulation of a sufficient number of T cells able to recognize the antigen. These specific T cells must then arrive and selectively accumulate at a site where the antigen is present. Since this process takes much longer than reactions involving antibodies already present throughout blood and tissue fluids, type IV reactions were first distinguished by their delayed onset and are still frequently referred to as delayed hypersensitivity reactions. They not only come on slowly, but, depending upon whether the antigen persists or is removed, they can be prolonged or relatively transient.

The T cells involved in type IV reactions are memory cells derived from prior stimulation by the same antigen. These cells persist for many months or years, so that persons who have become hypersensitive to an antigen tend to remain so. When restimulated by this antigen presented on the surface of the macrophages (or, sometimes, on other cells that can express MHC II antigens), the T cells secrete lymphokines, thereby triggering the cell-mediated immune response described earlier. Two common examples of delayed hypersensitivity—tuberculin-type and contact hypersensitivity—illustrate the various consequences of type IV reactions.

Tuberculin-type hypersensitivity. The familiar tuberculin test, used to determine whether a person has immunity to tuberculosis, is based on a delayed hypersensitivity reaction. In this test, small amounts of protein extracted from the mycobacteria that cause tuberculosis are injected into the skin. If reactive T cells are present—*i.e.,* the test is positive—redness and swelling appear at the injection site the next day, increase for a further day, and then gradually fade away. If examined histologically, the site of a positive reaction shows infiltration by lymphocytes and monocytes, increased fluid between the fibrous structures of the skin, and some cell death. If the reaction is severer and prolonged, some of the activated monocytes—which turn into macrophages when they leave the circulation—will have fused together to form large cells containing several nuclei. An accumulation of activated macrophages of this sort is termed a granuloma. Immunity to a number of other diseases (for example, leprosy, leishmaniasis, coccidiosis, and brucellosis) can also be gauged by the presence or absence of a delayed reaction to a test injection of the appropriate antigen. In all these cases, the test antigen provokes only a transitory response when the test is positive (and, of course, no response at all when the test is negative).

Formation of granulomas

The same cell-mediated mechanisms are invoked during actual infection by the living microbes, but in this case the response continues and the ensuing tissue damage and granuloma formation can cause serious damage. Moreover, in an actual infection, the microbes are often present inside the macrophages and are not necessarily localized in the skin. Large granulomas develop when the stimulus causing activation persists, especially if undegradable particulate materials are present and several macrophages, all attempting to ingest the same material, have fused their cell membranes to one another. The macrophages continue to secrete enzymes capable of breaking down proteins, and the normal structure of tissues in their neighbourhood becomes distorted. Although granuloma formation may be an effective way of isolating indigestible materials (whether or not of microbial origin) from the rest of the body, the damage caused by this immune mechanism can become much more serious than any damage caused by the infectious organisms. This is the case in such diseases as pulmonary tuberculosis, schistosomiasis, and certain fungal infections that become established within the body tissues rather than at their surface.

Contact hypersensitivity and dermatitis. A variety of chemicals can combine with the body's proteins so that these now have added antigenic determinants or are altered in some other way that makes them "foreign." Penicillin has already been mentioned in this connection, but the list includes a variety of other drugs as well as polyphenolic substances from plants (*e.g.,* poison ivy), hair dyes, industrial chemicals, and metals (such as chromium, nickel, or mercury when these have been converted into reactive forms by bacterial action or other means). When these substances are applied to the skin, they combine partly with some of the proteins of cells in the outer horny layer and partly with cells and proteins in the underlying tissue. Some of the latter are carried to the draining lymph nodes, where they stimulate those T cells that recognize the foreign antigens to multiply. When these T cells leave the nodes and enter the bloodstream, they can travel back to the skin where the chemical was first applied and some remains in the local tissues. There they become reactivated, infiltrating the skin and inducing a chronic inflammation. This partly resembles a tuberculin response, but it has an added element due to the fact that any cytotoxic T cells recognizing the antigen can also find a target. The clinical result is contact dermatitis, which can persist for many days or weeks. Treatment is by local application of corticosteroids, which greatly diminish lymphocyte infiltration, and by avoidance of further contact with the sensitizing agent.

Chemical alteration of body cells

Although delayed hypersensitivity can be a nuisance when it produces skin allergies, it makes up an important part of the immune defense against intracellular parasites (provided that the reaction succeeds in eliminating these), and it probably also plays a role in the containment of some tumours.

AUTOIMMUNE DISORDERS

In describing the nature and origin of the antigen receptors of B and T cells, it was explained that their enormous diversity is generated by a random process that inevitably gives rise to receptors for self determinants. It was also explained that lymphocytes bearing such receptors are eliminated or rendered impotent by several different mechanisms, so that the immune system does not normally generate significant amounts of antibodies or T cells reactive with the body's own components. Nevertheless, autoimmunity can occur, and some of the ways in which this causes damage have been mentioned in the preceding section on *Allergies.*

Basic processes that underlie autoimmunity. Immunologists cannot always explain why the mechanisms that normally prevent the development of autoimmunity have failed in a particular autoimmune disorder. They have, however, advanced a number of explanations for such failures.

Alteration of self antigens. Various mechanisms can result in self components being altered in some way so that they are perceived as foreign by the immune system. This can result from the attachment of new antigenic determinants as described under *Allergies.* It can also occur if the self components change part of their shape, so that previously unavailable helper T cells are stimulated and can cooperate with preexisting B cells to secrete autoantibodies. This can readily be shown to occur in experimental animals and is the most probable explanation of the production of the rheumatoid factors that cause the damage characteristic of rheumatoid arthritis (see below). Alterations caused by infectious organisms may explain why viral infection of specialized cells—such as those in the pancreas that secrete insulin or those in the thyroid gland that make thyroid hormone—often precedes the development of autoantibodies against the cells themselves and against their hormonal products.

Body cells altered by infection

Release of sequestered self antigens. Intracellular antigens and antigens found on tissues that are not in contact with the circulation are normally effectively segregated from the immune system. These may be regarded as foreign if they are released into the circulation as a result

of tissue destruction caused by trauma or infection. After sudden damage to the heart, for example, antibodies against heart muscle membranes regularly appear in the blood. They are usually transient and do no harm.

Cross-reaction with foreign antigens. This mechanism comes into play when an infective agent produces antigens so similar to those on normal tissue cells that antibodies reactive with these cells are stimulated. The streptococci that cause rheumatic fever make antigens cross-reactive with those on heart muscle membranes, and the trypanosomes that cause Chagas' disease in South America make antigens cross-reactive with the specialized nerve cells that regulate the orderly contraction of muscles in the bowel. Autoantibodies stimulated by external antigens in this way tend to persist and can cause serious damage.

Heredity. Several autoimmune diseases clearly run in families. Careful studies (for example, those comparing the incidence in identical twins with that in fraternal twins) have shown that the increased incidence of such autoimmune diseases cannot be explained by environmental factors. Rather, it stems from a hereditary (*i.e.,* genetic) tendency to contract the disease. Such disorders include thyroiditis (Graves' disease, myxedema, Hashimoto's disease), atrophic gastritis (including pernicious anemia), juvenile-type diabetes mellitus, and adrenal disease (Addison's disease). They are more common in persons who bear particular MHC I and MHC II antigens. The possession of these antigens does not imply that a person will contract such diseases, only that he is more liable to do so. Other diseases, such as rheumatoid arthritis or systemic lupus erythematosus, are associated with other combinations of MHC antigens. Researchers generally agree that the interaction of more than one gene, and perhaps several, is needed before a person develops such autoimmune diseases, but these genes have not been identified.

Inherited autoimmune diseases

Production of monoclonal autoantibodies. If a B cell capable of making autoantibody undergoes uncontrolled proliferation (similar to that which produces myelomas and lymphomas), the result is a clone of cells that produce self-reactive immunoglobulin. Although this malfunction of immunity is not common, some cases of autoimmune hemolytic anemia result from monoclonal antibodies against red blood cells. Other rare individuals produce monoclonal antibodies against normal plasma proteins, especially IgG.

Malfunction of suppressor T cells. The most general hypothesis regarding autoimmune disorders is that the normal regulatory mechanism effected by specific suppressor T cells has failed. This must indeed be true. But since the reason for the failure is not known, the hypothesis does no more than suggest that any treatment, apart from alleviating symptoms, should be designed to hold helper T cells in check and to promote suppressor T cells.

Examples of autoimmune disorders. The spectrum of autoimmune diseases is wide, ranging from those that involve a single organ to others that affect several different organs as a secondary consequence of the presence of immune complexes in the circulation. It is not possible in this article to discuss them all. The following are chosen to illustrate some of the very different complications that can arise.

Thyroiditis. The thyroid gland is an endocrine organ composed of closed sacs (follicles) lined with specialized cells. These cells secrete thyroglobulin, a large protein molecule containing iodine, into the interior of the follicles. Thyroglobulin acts as a storage form from which thyroid hormone is formed by degradation and released into the blood. The rate at which this occurs is regulated by thyroid stimulating hormone (TSH; a peptide made by the pituitary gland). TSH combines with receptors on the thyroid cells and stimulates them to make more thyroglobulin. In autoimmune thyroiditis, T cells occur that react with antigens in the thyroid cell membrane and damage the cells, causing the release of thyroglobulin. Antibodies form against thyroglobulin, against tiny projections (microvilli) on the thyroid cell surface, and against the receptors for TSH. These last, according to how they react with the receptors, either block them, rendering the gland unresponsive to TSH and very poorly functional, or they mimic the action of TSH and override the normal control by the pituitary. So, paradoxically, autoimmune thyroiditis can cause either thyroid deficiency or overactivity (Graves' disease). In either case the gland becomes infiltrated with lymphocytes and B cells and is partly destroyed. It may become completely destroyed, leading to myxedema.

Autoimmune hemolytic anemia. This condition occurs when an autoimmune attack destroys a person's red blood cells. Antibodies against a subject's own red cells appear during infections with the bacterium *Mycoplasma pneumoniae,* a rather uncommon cause of pneumonia. In this case the autoantibodies are directed against certain antigens that are present on all red cells irrespective of their blood group, and they are probably induced by a similar antigen in the microbes themselves. Anti-red-cell antibodies with a different specificity are often produced in persons treated for several months with the drug alpha methyldopa; the reasons for this are unknown. Other drugs, such as quinine, sulfonamides, or even penicillin, very occasionally cause hemolytic anemia. In such cases it is thought that the drug itself becomes bound to a protein adsorbed into red cells; complexes of antibodies against the drug in turn bind complement, and this lyses the cells. Autoantibodies against the red cells may be made by aberrant clones of B cells.

Antibodies against red cells vary greatly in avidity; that is, in the strength with which they bind to antigen. Avidity differences are a feature of antibodies in general, but they become very obvious in the case of anti-red-cell antibodies. Those that bind weakly can cause clumping of red cells at room temperature or below, but they come off red cells at body temperature (37° C [98.6° F]). Only those antibodies that function at 37° C or warmer normally cause hemolytic anemia, though even the "cold" antibodies can cause damage by clumping red cells in the blood vessels of cold hands or feet or other extremities.

Pernicious anemia. Pernicious anemia stems from a failure to absorb vitamin B_{12} (cobalamin), which is necessary for the proper maturation of red blood cells. It is characteristically accompanied by failure to secrete hydrochloric acid in the stomach (achlorhydria) and is in fact a symptom of severe autoimmune gastritis. In this condition autoantibodies are made against intrinsic factor, a protein secreted by the parietal cells in the stomach lining. Vitamin B_{12} must form a complex with intrinsic factor before it can be absorbed into the small intestine; the autoantibodies inhibit this and so prevent the vitamin from being taken into the body. The autoantibodies also destroy the acid-secreting cells, leading to achlorhydria.

Rheumatoid arthritis. Rheumatoid arthritis is a chronic inflammatory disease that affects connective tissues throughout the body but particularly in the peripheral joints. This inflammatory process results from autoimmune reactions that involve one or more autoantibodies referred to collectively as rheumatoid factor. These autoantibodies react with the constant region of the H chains in IgG molecules—in other words, the rheumatoid factors are anti-IgG antibodies.

Rheumatoid arthritis as an autoallergic reaction

Rheumatoid factors made by B cells in the synovial lining of the joints form complexes with IgG molecules; these complexes then trigger a type III hypersensitivity reaction, activating complement and thereby leading to inflammation within the joint. Rheumatoid factors can also be released into the blood, where they form complexes with circulating IgG molecules. If these immune complexes are not effectively removed by the liver and spleen, they can produce systemic effects similar to those precipitated by serum sickness.

Systemic lupus erythematosus (SLE). Systemic lupus erythematosus is one of a spectrum of diseases in which autoantibodies are made against the nucleic acid and protein constituents of normal cell nuclei. (Other diseases of this type include scleroderma, which is associated with thickening and hardening of the skin, and Sjögren's syndrome, which affects the lachrymal and salivary glands.) Damage in SLE is not due to the antibodies attacking healthy cells, since the nucleus is not accessible to antibodies outside the cells. Instead, it results from immune

complexes that form after nuclear constituents have been released into the bloodstream during the normal course of cell death and renewal or as a result of inflammation. The kinds of damage are those described in the preceding section under type III hypersensitivity, but there is a special tendency for the immune complexes to settle in the kidney and cause Bright's disease (glomerulonephritis). SLE affects about four times as many women as men.

It is a striking fact that most of the autoantibodies that occur in SLE patients react with the corresponding tissues or components in all mammalian species examined. These antibodies must, therefore, be directed against structures that have important functions and that have been conserved during the process of evolution. What initiates the production of autoantibodies of this sort is not known. In certain inbred strains of mice, however, nearly all the adult animals spontaneously develop autoimmune diseases very similar to SLE in humans. Since in some of these strains all the animals were congenitally heavily infected with viruses that were present in the circulation, it seemed a reasonable hypothesis that chronic stimulation by the nucleic acid of the virus caused the autoantibody production. But in other autoimmune strains there was no such viral infection. Some strains were deficient in suppressor T cells; one produced much less of IL-2 than normal; in one strain females were earlier and more severely affected than males (as in humans), but in another strain male mice were affected and not females. No single or simple defect or genetic factor has been identified that would account for the development of autoimmunity in these mice and thereby provide a clue to the cause of the disease in humans.

Other autoimmune disorders. Mechanisms similar to those that produce autoimmune hemolytic anemia can result in the formation of antibodies against granulocytes and platelets, although autoimmune attacks against these blood cells occur more rarely. Antibodies against other kinds of cells occur in a number of autoimmune diseases and may be a main cause of the damage in them. This is the case in myasthenia gravis, a disease characterized by muscle weakness. In this illness, autoantibodies react against receptors on muscle cells. Normally these receptors respond to acetylcholine, a neurotransmitter released from nerve endings that stimulates the muscle to contract. The autoantibodies in myasthenia gravis block or destroy the receptors, and muscle contraction is much weakened.

Myasthenia gravis

A different example is provided by Goodpasture's syndrome, a disorder in which autoantibodies form against the basement membrane of the blood vessels in the kidney glomeruli and in the air sacs of the lung. These antibodies cause severe kidney damage and lung hemorrhage.

Treatment. The treatment of autoimmune diseases depends upon what tissues are affected. If the effects are limited to a single endocrine organ that fails to secrete its normal hormone, regular administration of the missing hormone is sufficient to maintain reasonable health. Overactivity of the thyroid gland can be controlled by surgical removal of part of the thyroid or by the administration of a radioactive form of iodine that becomes concentrated in the thyroid and selectively irradiates the active portions of the gland. This latter therapy can restore normal thyroid function, but it may eventually lead to a correctable thyroid deficiency. In pernicious anemia, treatment formerly involved the oral administration of the missing intrinsic factor or of liver extracts rich in vitamin B_{12}, but patients now receive regular injections of pure cobalamin. Physicians treat conditions such as SLE with immunosuppressive drugs and corticosteroids. The former prevent B-cell proliferation, thereby damping down antibody production, including that of autoantibodies. The corticosteroids help suppress the inflammatory process. These drugs must be used carefully so as to avoid causing an immune deficiency.

CANCERS OF THE LYMPHOCYTES

At any stage in its development, from stem cell to mature form, a lymphocyte may undergo malignant (cancerous) transformation. The transformed cell is no longer constrained by the processes that regulate normal development, and it proliferates to produce a large number of identical cells. These retain the characteristics of the transformed cell's particular developmental stage, but they spread and displace those other cells that continue to develop normally. What exactly causes the unrestrained proliferation is not known, but in several instances parts of one chromosome have become detached from their normal position and reattached (translocated) elsewhere. Researchers have found that the sites where the chromosome breaks are close to sites where oncogenes are situated. Oncogenes are segments of DNA that are found in normal cell nuclei (where they may play a role in regulating cell multiplication during fetal life) as well as in tumour-producing viruses. This suggests that the translocation has allowed normally silent oncogenes to become activated, leading to uncontrolled growth and exaggeration of some normal activities of the cells. For example, B-cell tumours that produce large amounts of a particular class of immunoglobulin have translocations involving the chromosomes on which the genes controlling the synthesis of that kind of immunoglobulin reside.

Tumours arising from lymphocytes are given various names: they are called leukemias if the cancer cells become present in large numbers in the blood, lymphomas if they are mainly concentrated in lymphoid tissues, and (as was mentioned earlier) myelomas if they are B-cell tumours that secrete large amounts of immunoglobulin. Both T and B cells have surface antigens that are characteristic of different stages in their life cycle, and antibodies have been prepared that identify these antigens. It is, therefore, easy to determine the nature of the tumour cells. This is important because the different types respond to different therapies and because the chances of effecting a cure vary from type to type. Most leukemias arise in childhood, and one of the most satisfying advances in the drug treatment of cancers has been that, instead of all dying, the great majority of children with leukemia nowadays survive. Similarly, most cases of Hodgkin's disease, a common type of lymphoma that strikes mainly adults, can be cured by drugs, radiation, or a combination of both. Myelomas mostly arise in older subjects. They grow fairly slowly and are usually diagnosed by virtue of the characteristic immunoglobulin they secrete, which may be produced in such large amounts as to cause secondary damage such as kidney failure.

Leukemia, lymphoma, and myeloma

Evolution of the immune system

DEVELOPMENT OF IMMUNITY IN MAJOR ANIMAL GROUPS

Because the immune system is composed of cells and tissues that do not lend themselves to fossilization, it is impossible to trace the evolution of immunity from the paleontological record. It is, however, possible to study the immune capacity of living animals and, based on the relative positions of these animals in the evolutionary tree, to extrapolate a reasonable evolutionary history of the immune system.

Immune capacity among invertebrates. Even such primitive creatures as sponges have means of distinguishing self components from nonself components. Sponges grow in colonies, and all of the individuals in any one colony are alike. If, in an experiment, two sponges from the same colony are lashed tightly together, their tissues eventually will merge to form one sponge. But if two sponges from different colonies are secured together in the same way, the tissues that touch will die. How exactly this occurs is not known, but it is not like the immune response described in this article.

Primitive immune systems

When tissue grafts are made in animals higher up the evolutionary tree—between individual annelid worms or starfish, for example—the grafted tissue is commonly invaded by phagocytic cells and cells resembling lymphocytes, and it is destroyed. Yet tissues grafted from one part of the body to another on the same individual adhere and heal readily and remain healthy. So it seems that something akin to cellular immunity is present at this level of evolution, but these creatures do not have lymphoid tissues nor do they make immunoglobulins as do verte-

brates. The circulatory fluids of some invertebrates, such as the king crab (an arthropod), have a protein not unlike one of the proteins found in the complement sequence, which can bind to foreign cells and lyse them. Nothing akin to the complete complement system of mammals is present, however.

Immune capacity among vertebrates. Recognizable lymphocytes and immunoglobulins appear only in the vertebrates. The most primitive living vertebrates—the jawless fish (hagfish and lampreys)—do not have lymphoid tissues corresponding to a spleen or a thymus, and their immune responses, although demonstrable, are very weak and sluggish. As one ascends the evolutionary tree to the cartilaginous fish (sharks and rays) and the bony fish, both a thymus and a spleen are present; and antibody responses can be elicited quite easily, though only antibodies of the IgM class are detectable. Fish lack specialized lymph nodes, but they do have clusters of lymphocytes in the gut that may serve an analogous purpose.

Complete immune systems

It is not until one reaches the level of the terrestrial vertebrates—amphibians, reptiles, birds, and mammals—that a complete immune system with thymus, spleen, bone marrow, and lymph nodes becomes evident and that both IgM and IgG antibodies are made. Antibodies of the IgA class are only found in birds and mammals, and IgE antibodies are confined to mammals. So it appears that the most primitive devices for producing specific, acquired immunity gradually diversified to meet the new environmental hazards as animals moved out of the sea onto the land.

The evolution of the complement system may have occurred rather faster than the immunoglobulin system. The jawless fishes have complement components corresponding only to the later acting (*i.e.,* cytolytic, or cell-killing) aspects of complement function, but all higher vertebrates have components similar to the complete system of mammals. The fact that the complement system has been so well conserved during evolution implies not only that it has been of great biological value, but also that throughout the evolution of the immune system in higher vertebrates, the interaction of complement and immunoglobulins has been part of their joint function.

GENETIC ORIGINS OF THE IMMUNE SYSTEM

Researchers have analyzed the exact structure not only of the antigen receptors of lymphocytes but also that of other receptors not concerned with specific antigen recognition (such as those that bind the constant "tail" part of H chains in immunoglobulins) and of the molecules that constitute the MHC class I and class II antigens. The striking fact has emerged that the general structure of all these molecules is unexpectedly similar. Each is composed of one or more subunits, or domains, of nearly the same size, formed into a loop by bonds between sulfur atoms on the amino acids at the ends. Although each domain is different and serves a different function in the molecule as a whole, the number and order of the amino acids forming each domain is far more similar than would be expected if each had arisen independently in the course of evolution. Equally remarkable is the fact that nerve cells and T lymphocytes in mice and rats carry a surface protein termed Thy 1, the function of which is unknown, that also has this same basic structure and a similar arrangement of amino acids. These similarities suggest that the genes for all of these molecules originated from some primitive gene involved in the recognition of one cell by another, which is required for orderly development of a complex, multicellular organism, and that during evolution they have acquired different special functions. The processes whereby one ancestral gene could have given rise to such a family of genes include gene duplication, crossing over, and mutation, all of which are discussed in detail in GENETICS AND HEREDITY, THE PRINCIPLES OF.

Computing the age of evolving immune systems

Not surprisingly, molecules that have a similar function in different species (*e.g.,* immunoglobulins or MHC components) show an even closer resemblance. By analyzing the number of differences in the amino acids and their position in the polypeptide chains that make up IgM and IgG molecules in, for example, humans, mice, and rabbits and by making reasonable assumptions about mutation rates, scientists can estimate roughly how many generations would have had to elapse—and hence how much time—for the present immunoglobulins of these species to have evolved from a common ancestral IgM-like molecule. Such calculations suggest that divergence from the ancestral immunoglobulin took place some 200,000,000 years ago. Based on paleontological evidence, this is about the same as the time at which amphibians are thought to have diverged from the main vertebrate line. So one may conclude that a functional immune system arose even earlier and has continued to serve its purpose ever since.

BIBLIOGRAPHY. Many textbooks of immunology, all of which aim to be comprehensive, have appeared in recent years, and all have helpful diagrams. The following are listed in order of increasing complexity: J.H.L. PLAYFAIR, *Immunology at a Glance,* 3rd ed. (1984); IVAN M. ROITT, JONATHAN BROSTOFF, and DAVID K. MALE, *Immunology* (1985); JOSEPH A. BELLANTI (ed.), *Immunology: Basic Processes,* 2nd ed. (1985); and EMIL R. UNANUE and BARUJ BENACERRAF, *Textbook of Immunology,* 2nd ed. (1984). JAN KLEIN, *Immunology, the Science of Self-Nonself Discrimination* (1982), explains clearly how and why principles were discovered. For the history of immunology, see J.H. HUMPHREY and R.G. WHITE, *Immunology for Students of Medicine,* 3rd ed. (1970), especially the introductory chapter; and J.H. PARISH, *Victory with Vaccines: The Story of Immunization* (1968). Evolution of the immune system is well described in ELWIN L. COOPER, *General Immunology* (1982). For genetic aspects, see H. HUGH FUDENBERG *et al., Basic Immunogenetics,* 3rd ed. (1984). At a more advanced level, clinical aspects are covered by P.J. LACHMAN and D.K. PETERS (eds.), *Clinical Aspects of Immunology,* 4th ed., 2 vol. (1982); MAX SAMTER (ed.), *Immunological Diseases,* 3rd ed., 2 vol. (1978); and G.L. ASHERSON and A.D.B. WEBSTER, *Diagnosis and Treatment of Immunodeficiency Disease* (1980). For immunochemistry, see ELVIN A. KABAT, *Structural Concepts in Immunology and Immunochemistry,* 2nd ed. (1976); and L.E. GLYNN and M.W. STEWARD, *Immunochemistry: An Advanced Textbook* (1977). A good simple account of monoclonal antibodies is given in KAROL SIKORA and HOWARD M. SMEDLEY, *Monoclonal Antibodies* (1984).

(J.H.Hy.)